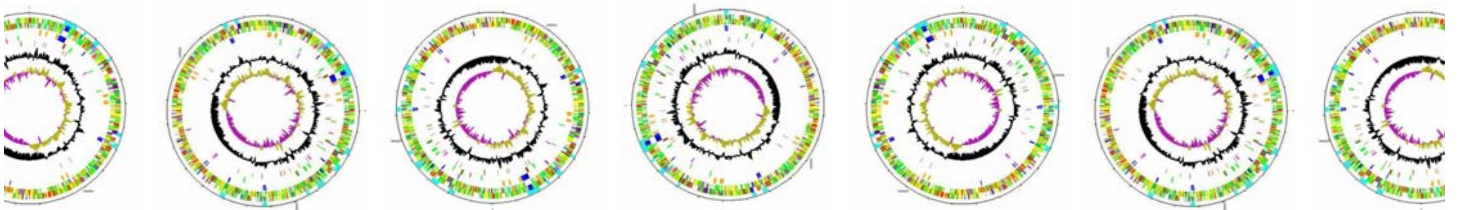


# Workshop on Bacterial Genomics

28-30 September 2005

*Held at:*

Ciutat de les Arts i de les Ciències - Valencia (Spain)



## Timetable

### Wednesday, 28th September

08:30-09:30 Registration  
09:30-10:00 Open ceremony. Andrés Moya.  
10:00-10:30 Coffee  
10:30-11:00 Introduction. Julian Parkhill  
11:00-14:00 Artemis: Guided exercises. Nicholas Thomson.  
14:00-15:00 Lunch  
15:00-16:00 Gene Prediction. Nicholas Thomson.  
16:00-17:00 ACT: Guided exercises. Ana Cerdeño.  
17:00-17:30 Coffee  
17:30-19:30 ACT: Guided exercises (cont'd). Ana Cerdeño.

19:30-20:30. **Public conference. Julian Parkhill**

21:00 Official dinner. L'Oceanogràfic. Submarine Restaurant.

### Thursday, 29th September

09:00-10:30 ACT: Guided exercises (cont'd). Ana Cerdeño  
10:30-11:00 Coffee  
11:00-12:30 Generating ACT comparison files. Ana Cerdeño.  
12:30-14:00 Jemboss + Internet Genome Resources. Tim Carver.  
14:00-15:00 Lunch  
15:00-16:30 Internet Genome Resources. Ana Cerdeño.  
16:30-17:00 Coffee  
17:00-19:00 Data mining using GeneDB. Christopher Peacock.

19:30-20:30. **Public conference. Jean Marie Claverie.**

Visit to two Scientific Centers in Valencia

19:30-22:30. Visit to CSAT and IVI.

Science Bar

21:00-22:15. Museum's Bar.

### Friday, 30th September

09:00-10:30 Comparative genomics. Francisco Silva and Amparo Latorre  
10:30-11:00 Coffee  
11:00-12:30 Comparative genomics (cont'd). Francisco Silva and Amparo Latorre.  
12:30-14:00 Phylogenomics. Fernando González and Rosario Gil.  
14:00-15:00 Lunch  
15:00-16:30 Phylogenomics (cont'd). Fernando González and Rosario Gil  
16:30-17:30 Genome flexibility. Alex Mira.  
17:30-19:30 Annotation summary exercise or own seq + Mop-up.

## Glossary of Abbreviations and Terms

<b>ACT</b>	Artemis Comparison Tool.
<b>BLAST</b>	Basic local alignment search tool.
<b>CDS</b>	Coding sequence (Gene with no biological evidence for expression).
<b>CNRS</b>	Centre National De La Recherche Scientifique.
<b>DDBJ</b>	DNA data bank of Japan.
<b>EBI</b>	European Bioinformatics Institute, Hinxton. An outstation of the European Molecular Biology Laboratory.
<b>EMBL</b>	European Molecular Biology Laboratory, the name of the European DNA database.
<b>EST</b>	Expressed sequence tag.
<b>Fasta</b>	Part of the 'Fast' repertoire of global alignment search tools.
<b>Flatfile</b>	A simple text file used as an alternative to a database to storing data.
<b>GENE-IT</b>	Is a company that collaborates with the EBI and others to discover the functions of genes through comparative genomics.
<b>HMM</b>	Hidden Markov Model.
<b>INRA</b>	Institut National De La Recherche Agronomique.
<b>InterPro</b>	A search tool which brings together many of the commonly used signature databases for sequence searching.
<b>LINUX</b>	A packaged version of UNIX for the PC.
<b>mRNA</b>	messenger RNA, processed RNA molecule to be translated to form protein.
<b>NCBI</b>	National Centre for Biotechnology Information. Part of the U.S. National Library of Medicine (NLM), National Institutes of Health (NIH).
<b>PRINTS</b>	Proteins Finger print database, a compendium of protein finger prints.
<b>PFAM</b>	Protein family, a searchable database of protein domains.
<b>ProDom</b>	a comprehensive set of protein domain families.
<b>Prosite</b>	Database of protein families and domains.
<b>PSU</b>	Pathogen Sequencing Unit.
<b>RFAM</b>	A searchable database of RNA families.
<b>SIB</b>	Swiss Institute of Bioinformatics (SIB).
<b>SignalP</b>	A program to predict the presence and location of signal peptide cleavage sites in amino acid sequences.
<b>SMART</b>	Simple Modular Architecture Tool.
<b>SWISS-PROT</b>	is a curated protein sequence database.
<b>TIGR</b>	The Institute of Genome Research.
<b>TIGRfam</b>	The Institute of Genome Research protein family database.
<b>TMHMM</b>	Program for prediction of transmembrane helices in proteins.
<b>TrEMBL</b>	Computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT.
<b>UNIX</b>	A computer operating system.

# Index

## Index

### Module 1: *Artemis*

Exercise 1

Exercise 2

### Module 2: *Comparative Genomics*

Exercise 1

Exercise 2

Exercise 3

Exercise 4

### Module 3: *Generating ACT comparison files using BLAST*

Exercise 1

Exercise 2

Exercise 3

Exercise 4

### Module 4: *Jembooss*

Exercise 1

Exercise 2

### Module 5: *Genome Resources*

Section 1

Exercise 1

Exercise 2

Exercise 3

Section 2

Exercise 1

Exercise 2

Section 3

### Module 6: *Data Mining using GeneDB*

## References

## Appendices



# Module 1

# Artemis

## Introduction

Artemis (Rutherford *et al.*, (2000) is a DNA viewer program, written by Kim Rutherford, and used for both Prokaryotic and Eukaryotic annotations. It allows the user to get away from the relatively faceless EMBL and Genbank style database files and view the sequence in a graphical and highly interactive format. Artemis is designed to present multiple lines of information within a single context. This manifests itself as being able to zoom in to look for fine DNA motifs as well as being able to zoom out and bring into view operons, several kilobases of a genome or in fact to view an entire genome in one screen. It is also possible to perform quite sophisticated analyses and store the output within the 'Artemis environment' to be accessed later.

## Aims

The aim of this Module is for you to become familiar with the basic functioning of Artemis by using a series of worked examples. These examples are designed to take you through the most immediately useful functions. However, there will be time, and encouragement, for you to explore other menus; nooks and crannies of Artemis that are not featured in the exercises in this manual. Like all the Modules in this workshop, the key is 'if you don't understand please ask'.

## Artemis Exercise 1 Part I

### 1. Starting up the Artemis software

Navigate your way into the correct directory for this module

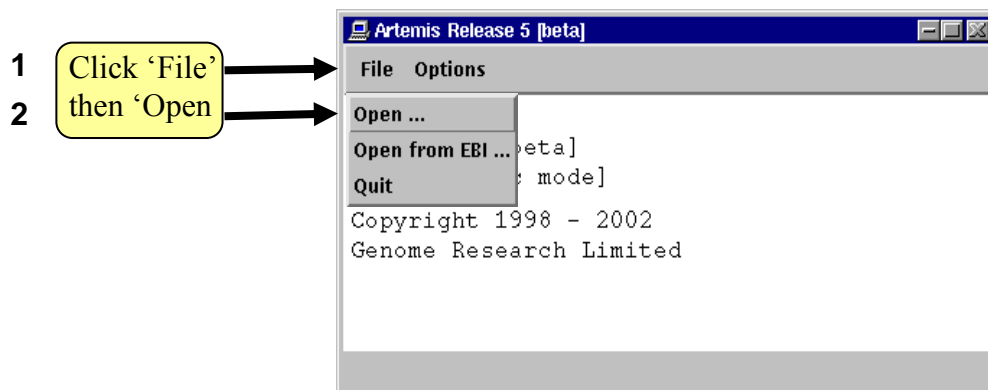
Then type:

art & [return]

A small start-up window will appear (see below).

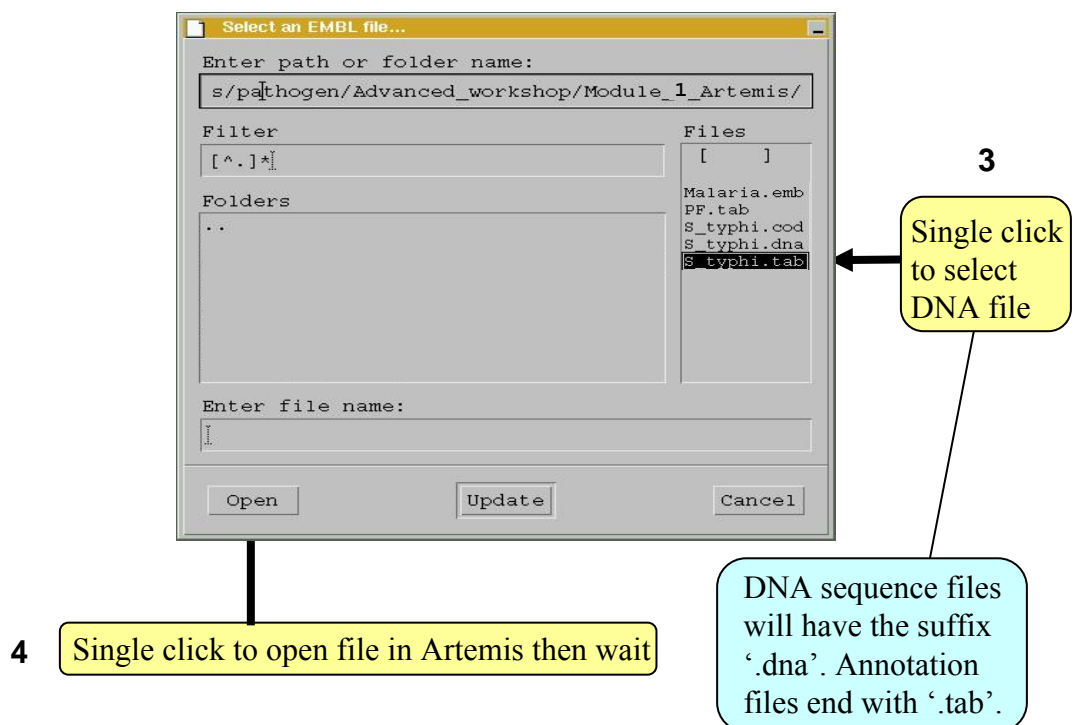
Now follow the sequence of numbers to load up the *Salmonella typhi* chromosome sequence.

Ask a demonstrator for help if you have any problems.



For simplicity it is a good idea to open a new start up window for each Artemis session and close down any sessions once you have finished an exercise.

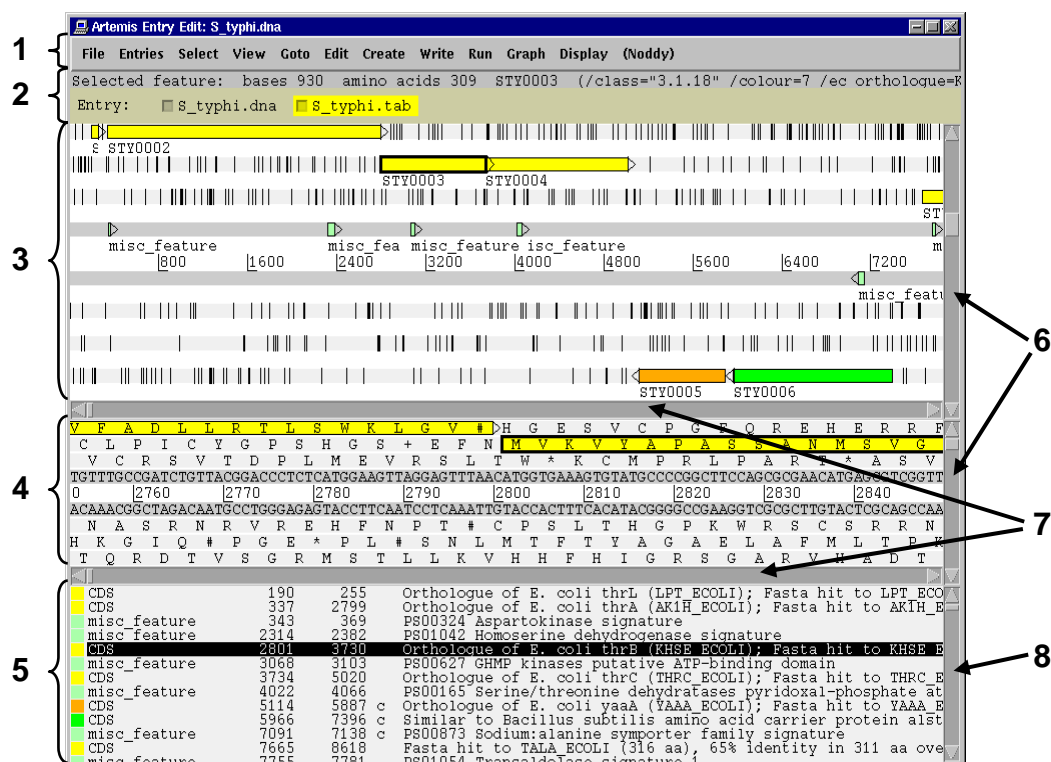
In the 'Options' menu you can switch between prokaryotic and eukaryotic mode.





### 3. The basics of Artemis

Now you have an Artemis window open let's look at what's in there.



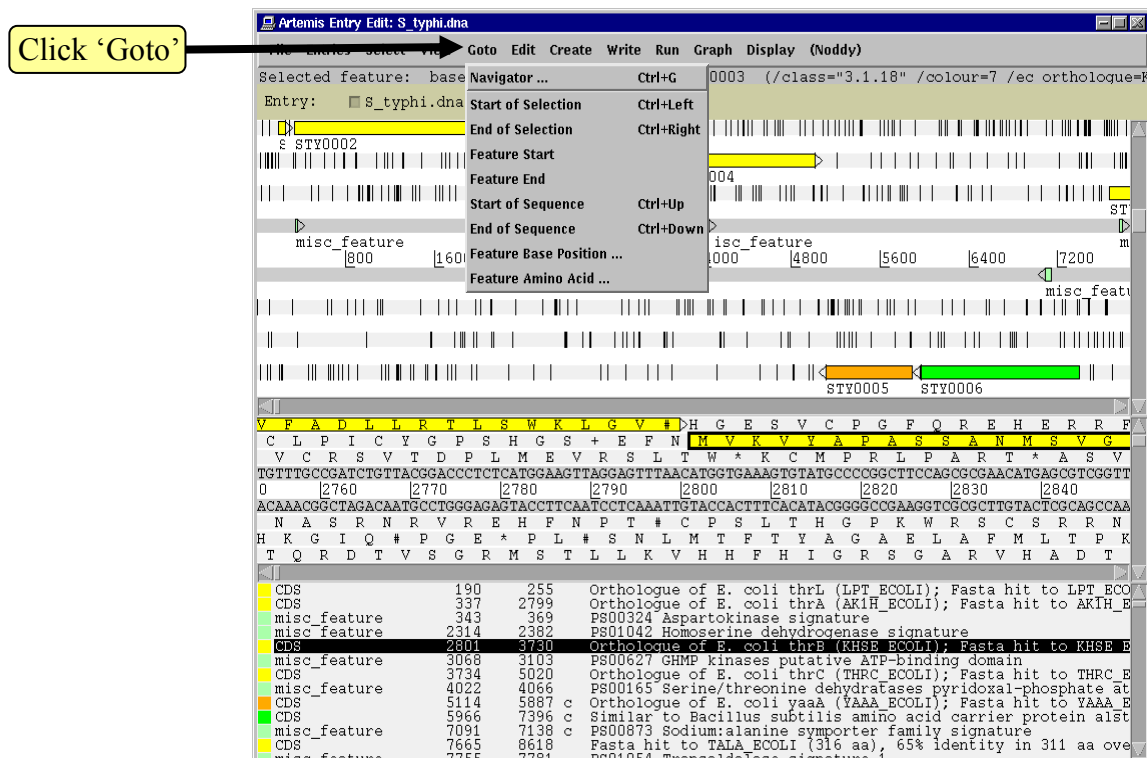
1. Drop-down menus. There's lots in there so don't worry about them right now.
2. Shows what entries are currently loaded (bottom line) and gives details regarding the feature selected in the window below; in this case gene STY0003 (top line).
3. This is the main sequence view panel. The central 2 grey lines represent the forward (top) and reverse (bottom) DNA strands. Above and below those are the 3 forward and 3 reverse reading frames. Stop codons are marked as black vertical bars. Genes and other features (eg. Pfam and Prosite matches) are displayed as coloured boxes. We will refer to genes as coding sequences or CDSs from now on.
4. This panel has a similar layout to the main panel but is zoomed in to show nucleotides and amino acids. Double click on a gene in the main view to see the zoomed view of the start of that gene. Note that both this and the main panel can be scrolled left and right (7, below) zoomed in and out (6, below).
5. This panel lists the various features in the order that they occur on the DNA with the selected gene highlighted. The list can be scrolled (8, below).
6. Sliders for zooming view panels.
7. Sliders for scrolling along the DNA.
8. Slider for scrolling feature list.

## 4. Getting around in Artemis

The 3 main ways of getting to where you want to be in Artemis are the Goto dropdown menu, the Navigator and the Feature Selector. The best method depends on what you're trying to do and knowing which one to use comes with practice.

### 4.1 The 'Goto' menu

The functions on this menu (ignore the Navigator for now) are shortcuts for getting to locations within a selected feature or for jumping to the start or end of the DNA sequence. This one's really intuitive so give it a try!



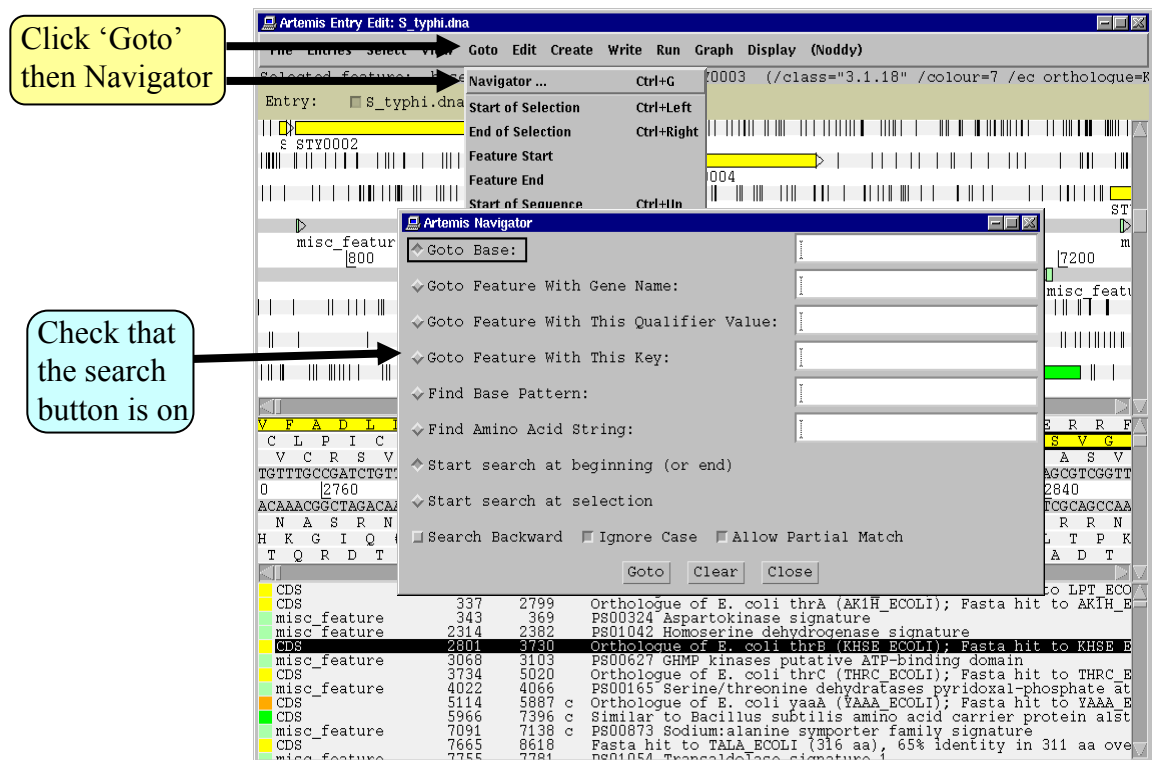
It may seem that 'Goto' 'Start of Selection' and 'Goto' 'Feature Start' do the same thing. Well they do if you have a feature selected but 'Goto' 'Start of Selection' will also work for a region which you have highlighted by click-dragging in the main window. So yes, give it a try!

#### Suggested tasks:

1. Zoom out, highlight a large region of sequence by clicking the left hand button and dragging the cursor then go to the start and end of the highlighted region.
2. Select a gene then go to the start and end.
3. Go to the start and end of the genome sequence.
4. Select a gene. Within it, go to a base (nucleotide) and/or amino acid of your choice.

## 4.2 Navigator

The Navigator panel is fairly intuitive so open it up and give it a try.



Suggestions of where to go:

1. Think of a number between 1 and 4809037 and go to that base (notice how the cursors on the horizontal sliders move with you).
2. Your favourite gene name (it may not be there so you could try 'fts').
3. Use 'Goto Feature With This Qualifier value' to search the contents of all qualifiers for a particular term. For example using the word 'pseudogene' will take you to the next feature with the word 'pseudogene' in any of its qualifiers. Note how repeated clicking of the 'Goto' button takes you through the pseudogenes as they occur on the chromosome.
4. tRNA genes. Type 'tRNA' in the 'Goto Feature With This Key'.
5. Regulator-binding DNA consensus sequence (real or made up!). Note that degenerate base values can be used (Appendix VIII).
6. Amino acid consensus sequences (real or made up!). You can use 'X's. Note that it searches all six reading frames regardless of whether the amino acids are encoded or not.

What are Keys and Qualifiers? See Appendix III

Clearly there are many more features of Artemis which we will not have time to explain in detail. Before getting on with this next section it might be worth browsing the menus. Hopefully you will find most of them easy to understand.

## Artemis Exercise 1 Part II

This part of the exercise uses the files and data you already have loaded into Artemis from Part I. By a method of your choice go to the region located between bases 2188349 to 2199512 on the DNA sequence. This region is bordered by the *fbaB* gene which codes for fructose-bisphosphate aldolase. You can use either the Navigator, Feature Selector or Goto functions discussed previously to get there. The region you arrive at should look similar to that shown below.

The screenshot shows the Artemis genome browser interface. The main window displays a DNA sequence with various features. The features are represented by colored bars and labels. The bottom panel shows a list of features with their coordinates and descriptions.

Feature Type	Start	End	Description
misc_feature	2188349	2199512	c Base composition: 37.8 % G+c
CDS	2188394	2189107	c Unknown function. Contains possible N-terminal signal sequen
CDS	2189209	2189652	c Unknown function. Contains probable N-terminal signal sequen
CDS	2189768	2190217	c Unknown function
CDS	2190285	2190764	c Unknown function. Contains possible N-terminal signal sequen
RBS	2190771	2190775	c possible RBS
CDS	2190874	2191476	c Unknown function. Contains possible N-terminal signal sequen
CDS	2191545	2191823	c Unknown function
CDS	2191793	2192488	c Unknown function
CDS	2192559	2193059	c Similar to Neisseria meningitidis hypothetical protein NMB04



Once you have found this region have a look at some of the information that is available to you:-

Information to view:

### **Annotation**

If you click on a particular feature you can view the annotation attached to it: select a CDS feature (or any other feature) and click on the Edit menu and select Edit Selected Feature. A window will appear containing all the annotation that is associated with that CDS. The format for this information is constrained by that which can be submitted to the EMBL database as seen in Module 1.

### **Viewing amino acid or protein sequence**

Click on the view menu and you will see various options for viewing the bases or amino acids of the feature you have selected, in two formats i.e. EMBL or FASTA. This can be very useful when using other programs that are not integrated into Artemis e.g. those available on the Web that require you to cut and paste sequence into them.

### **Plots/Graphs**

Feature plots can be displayed by selecting a CDS feature then clicking 'View' and 'Show Feature Plots'. The window which appears shows plots predicting hydrophobicity, hydrophilicity and coiled-coil regions for the protein product of the selected CDS.

### **Load additional files**

The results from Prosite searches run on the translation of each CDS should already be on display as pale-green boxes on the grey DNA lines. The results from the Pfam protein motif searches are not shown, but can be viewed by loading the appropriate file. Click on 'File' then 'Read an Entry' and select the file PF.tab. Each Pfam match will appear as a coloured blue feature in the main display panel on the grey DNA lines. To see the details click the feature then click 'View' then 'View Selection' or click 'Edit' then Edit Selected Features'. Please ask if you are unsure about Prosite and Pfam.

### **Viewing the results of database searches**

Click the 'View' menu, then select 'Search Results' and then 'Fasta results'. The results of the database search will appear in a scrollable window. If you click on the button at the bottom of this window labelled 'view in browser', then the results will be posted into an internet browser window. Within this window there are many active links (coloured blue), to external sources of information such as the original database entries for all those aligning to your sequence, as well as information stored in PubMed, PFAM and many others. Have a play.

Further information on specific Prosite or Pfam entries can be found on the web at

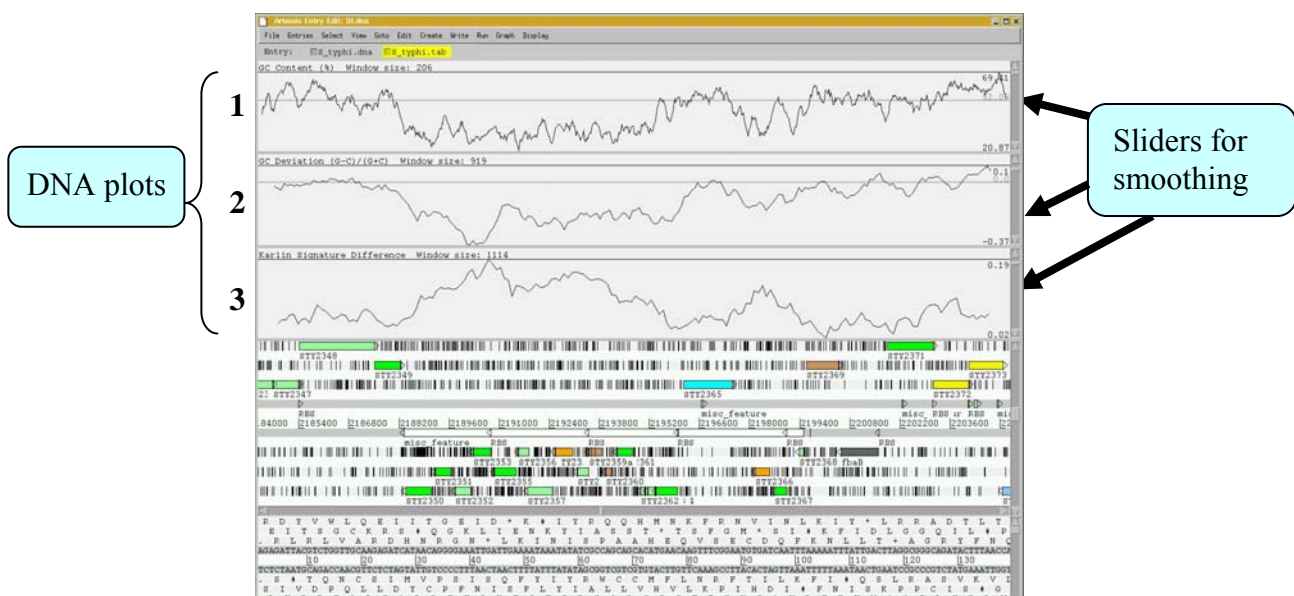
<http://ca.expasy.org/prosite> and <http://www.sanger.ac.uk/software/Pfam/tsearch.shtml>



In addition to looking at the fine detail of the annotated features it is also possible to look at the characteristics of the DNA covering the region displayed. This can be done by adding in to the display various plots showing different characteristics of the DNA. This information is generated dynamically by Artemis and although this is a relatively speedy exercise for a small region of DNA, on a whole genome view (we will move onto this later) this may take a little time so be patient.

**To view the graphs:**

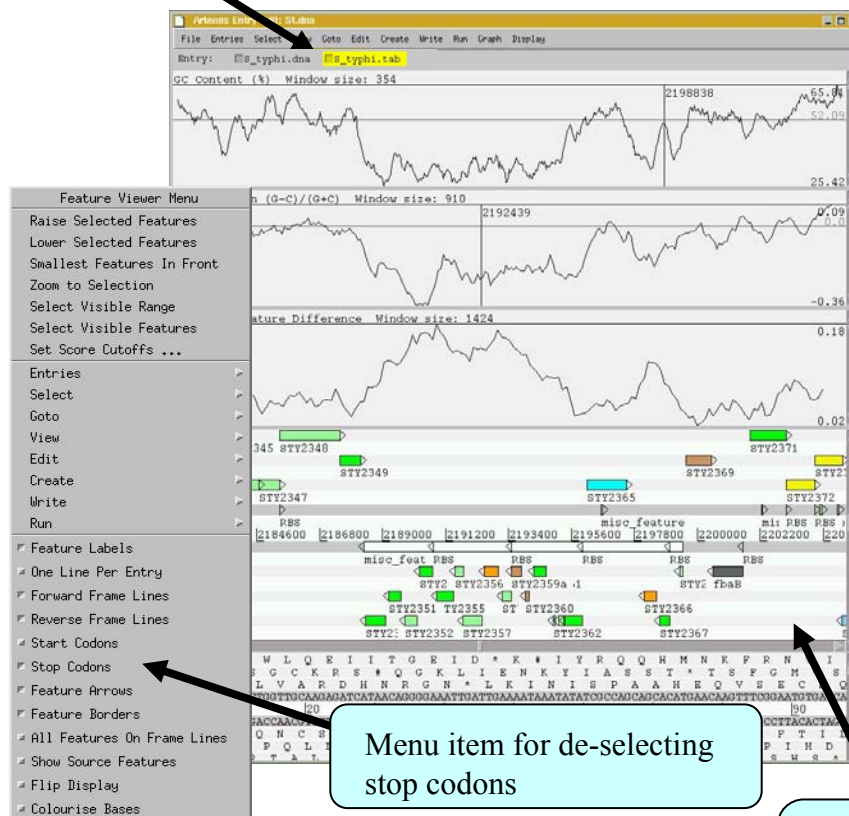
Click on the ‘Graph’ menu to see all those available. Perhaps some of the most useful plots are the ‘GC Content (%)’ **(1)** ‘GC Deviation’ **(2)** and ‘Karlin signature plots’ **(3)** as shown below. To adjust the smoothing of the graph you change the window size over which the points on the graph are calculated, using the sliders shown below. If you are not familiar with any of these please ask.



Notice how several of the plots show a marked deviation around the region you are currently looking at. To fully appreciate how anomalous this region is move the genome view by scrolling to the left and right of this region. The apparent unusual nucleotide content of this region is indicative of laterally acquired DNA that has inserted into the genome.

As well as looking at the characteristics of small regions of the genome, it is possible to zoom out and look at the characteristics of the genome as a whole. To view the entire genome use the sliders indicated below. However, be careful zooming out quickly with all the features being displayed, as this may temporarily lock up the computer. To make this process faster, and clearer, switch off stop codons by clicking with the right mouse button in the main view panel. A menu will appear with an option to de-select stop codons (see below). If you have any problems ask a demonstrator.

To de-select the annotation click here.



No stop codons shown on frame lines

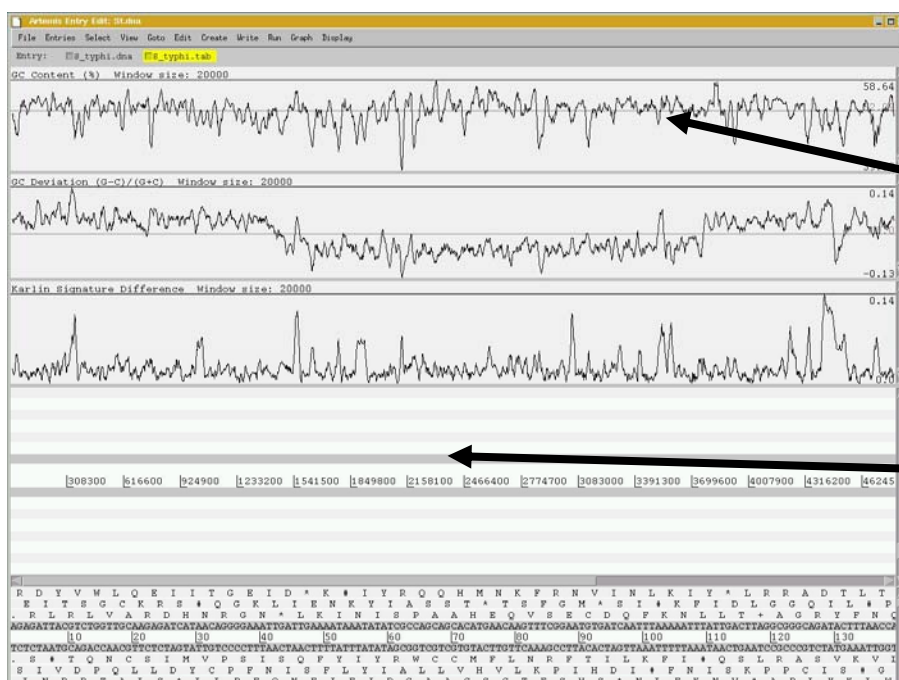
You will also need to temporarily remove all of the annotated features from the Artemis display window. In fact if you leave them on, which you can, they would be too small to see when you zoomed out to display the entire genome. To remove the annotation click on the S\_typhi.tab entry button on the grey entry line of the Artemis window shown above.

Your Artemis window should now look similar to the one shown below.



One final tip is to adjust the scaling for each graph displayed before zooming out. This increases the maximum window size over which a single point for each plot is calculated. To adjust the scaling click with the right mouse button over a particular graph window. A menu will appear with a series of values for the maximum window size (see above), select 20000. You should do this for each graph displayed.

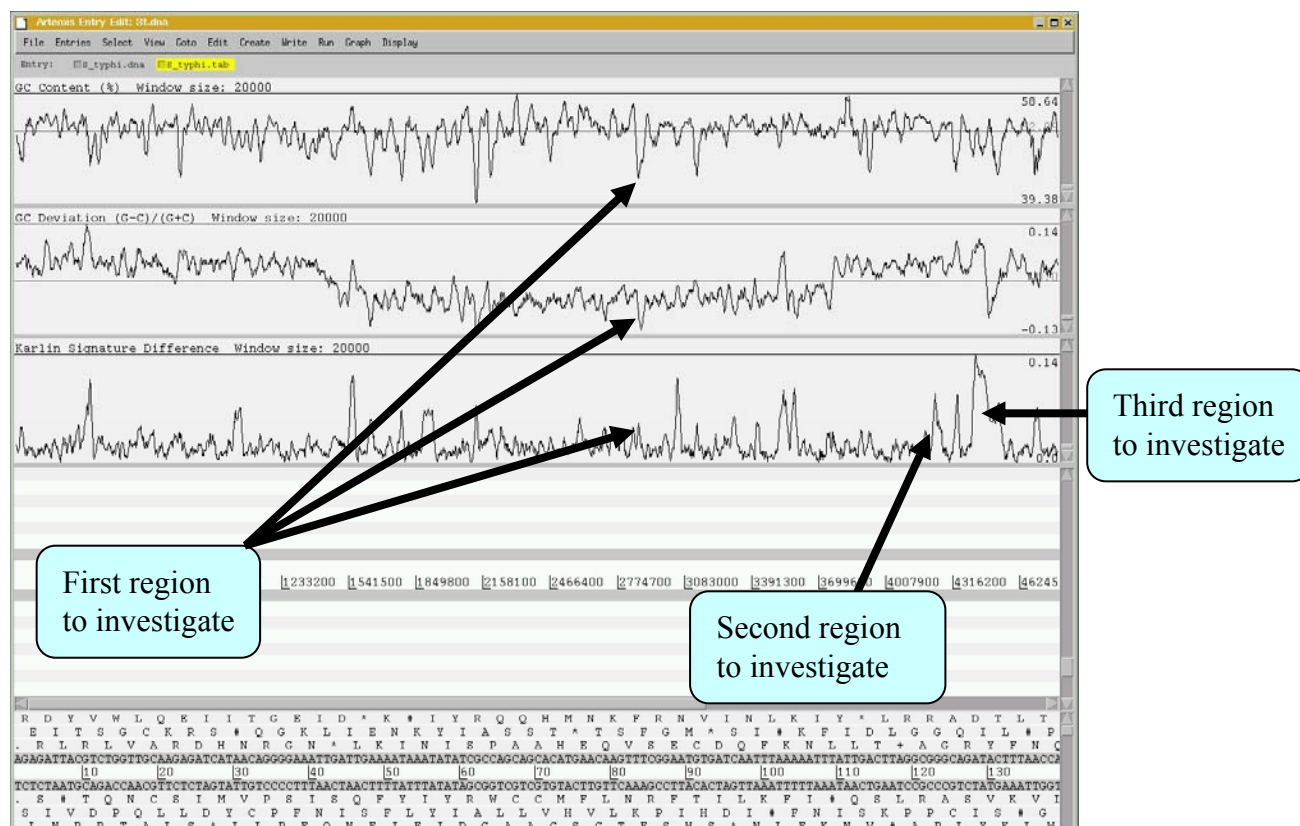
You are now ready to zoom out by dragging or clicking the slider indicated above. Once you have zoomed out fully to see the entire genome you will need to adjust the smoothing of the graphs using the vertical graph sliders as before to have a similar view to that shown below.



Click with the left mouse button in a graph window. A line and a number will appear. The number is the relative position within the genome (bps).

Click and drag to highlight a region on the main DNA line. Notice that the boundaries of this region should now be marked in the graph windows that your previously clicked in.

## Artemis Exercise 1 Part III



There are many examples where these anomalous regions of DNA within a genome have been shown to carry laterally acquired DNA. In this part of the exercise we are going to look at several of these regions in more detail. Starting with the whole genome view, note down the approximate positions and characteristics of the three regions shown above. Remember the locations of the peaks are given in the graph window if you click with the left mouse button within it.

Genome location	Characteristics of DNA plots
<b>Region 1 :</b> 2,860,000 bps	peak - karlin, troughs for G+C and CG deviation
<b>Region 2 :</b>	
<b>Region 3 :</b>	

We will now zoom back into the genome to look in more detail at the first of these three peaks. Zoom into this position by first clicking on the DNA line at approximately the correct location. If you then use the vertical side slider to zoom back in, Artemis will go to the location you selected. Remember that in order to see the CDS features lying within this region you will need to turn the annotation (S\_typhi.tab) entry back on.

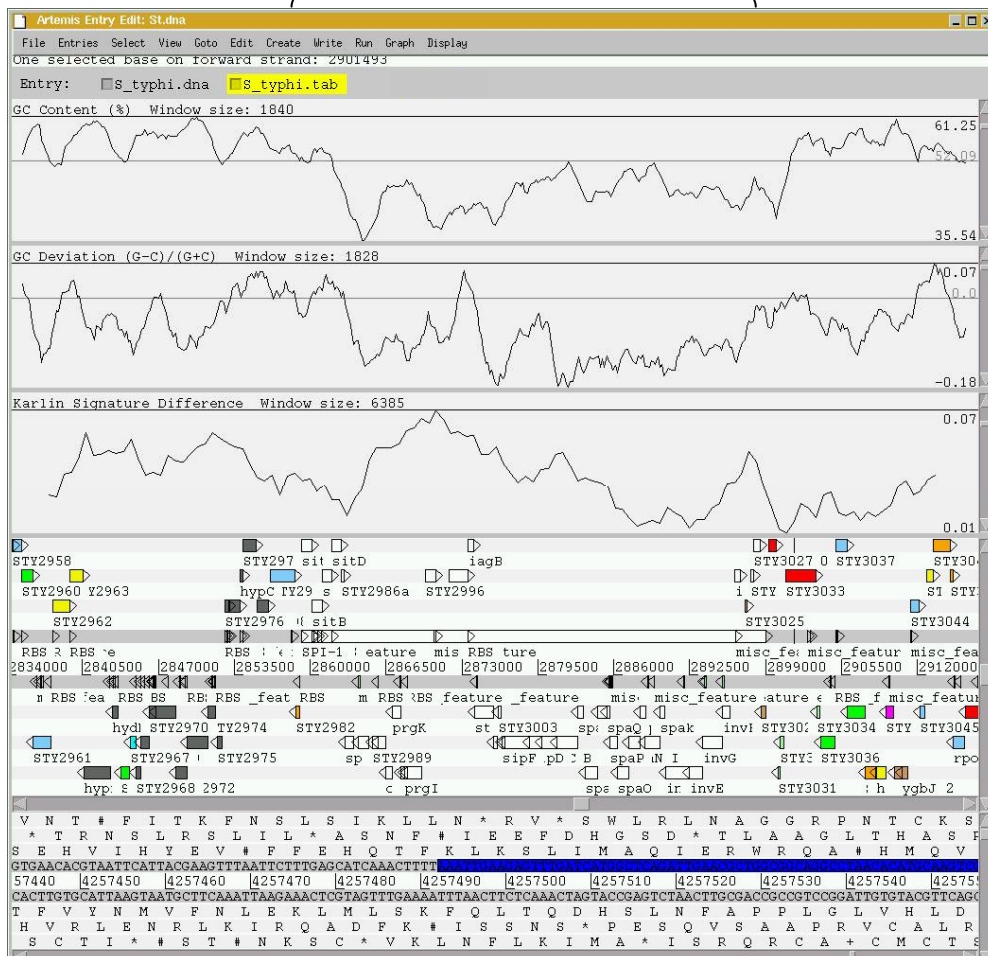


The region you should be looking at is shown below and is a classical example of what is referred to as a *Salmonella* pathogenicity island (SPI). The definitions of what actually constitutes a pathogenicity island are quite diverse. However, below is a list of characteristics which are commonly seen within these regions, as described by Hacker et al., 1997.

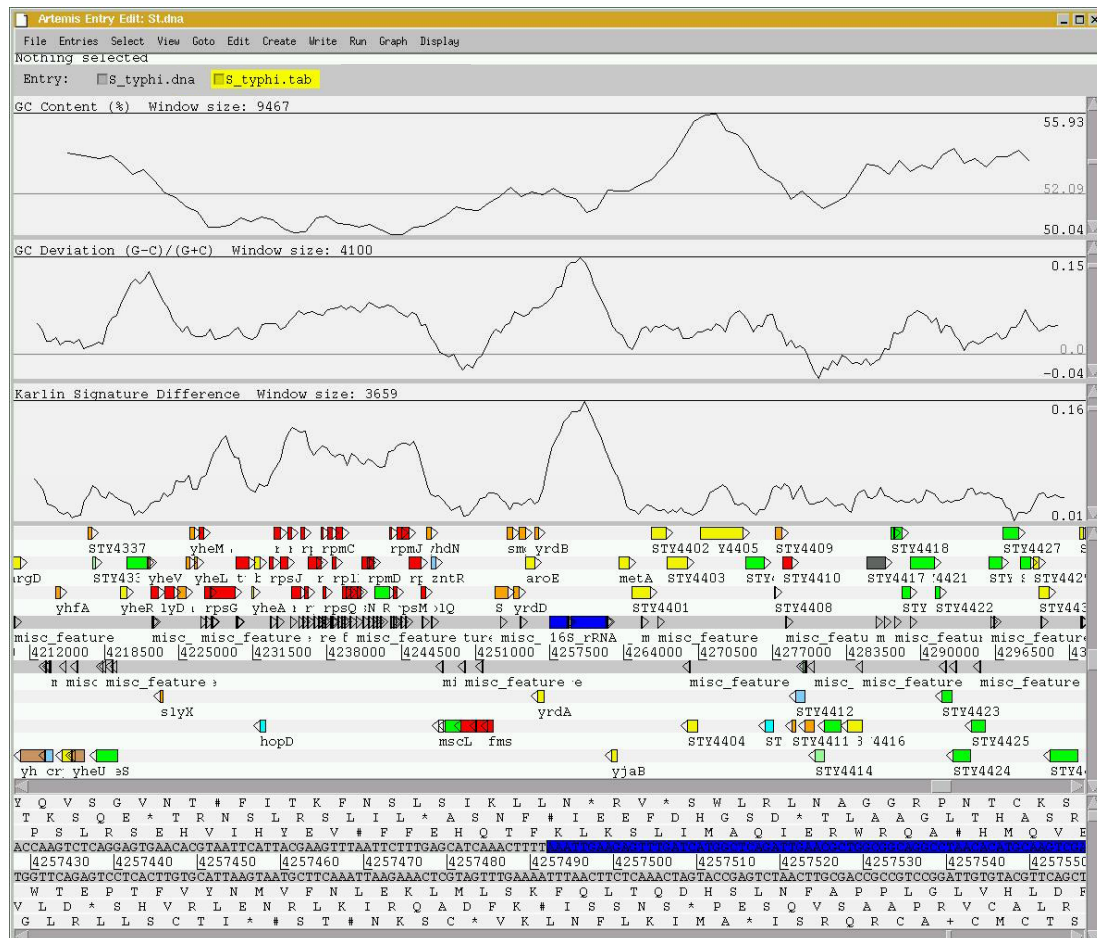
1. Often inserted alongside stable RNA's
2. Atypical G+C contents.
3. Carry virulence-related functions
4. Often carry genes encoding transposase or integrase-like proteins
5. Unstable and self-mobilisable
6. Of limited phylogenetic distribution

Have a look in and around this region and look for some of these features.

### Region 1 SPI-1



## Region 2



Use one of the methods you have already used to take you to the second region of interest that you noted down.

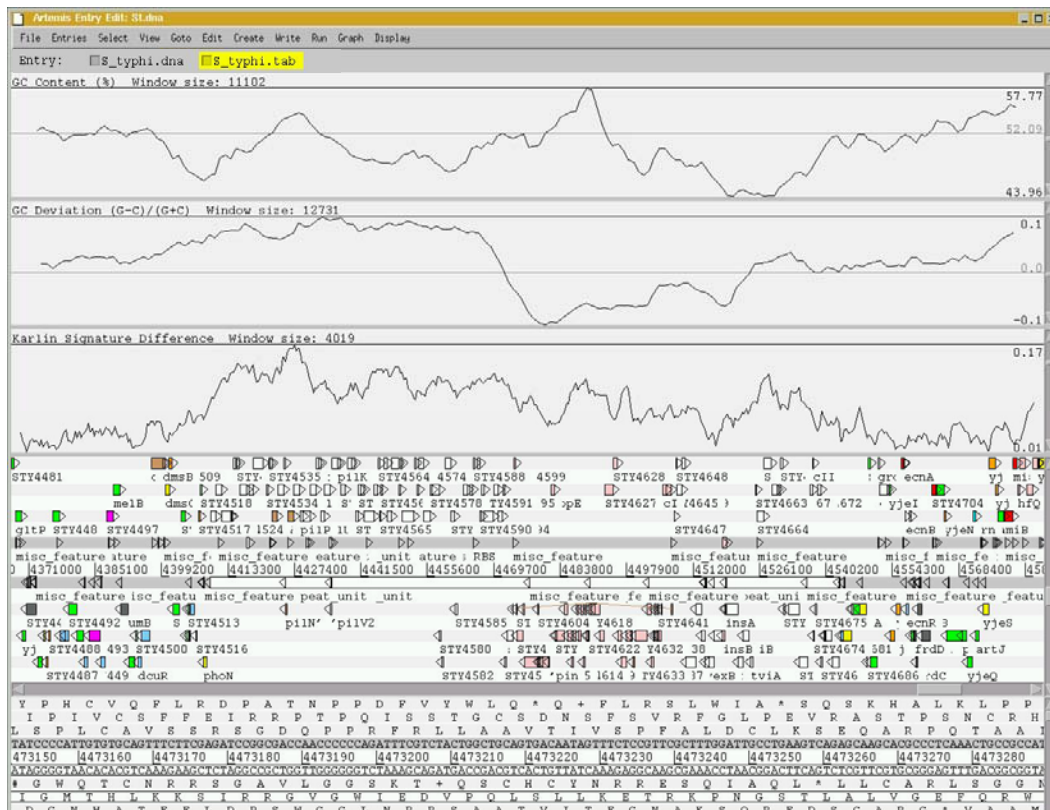
Region two acts as a cautionary note when looking at anomalous regions within a genome. Have a look at the CDSs within this region.

Does this region:

- have any of the characteristics of pathogenicity island
- are the genes within this region essential or dispensable.

Is it possible that the atypical base composition of this region is not a consequence of having originated from a foreign host. The base composition may actually be reflective of the tight sequence constraints under which this region has been maintained, in contrast to the background level sequence variation in the rest of the genome.

## Region 3



Go to region 3 as before.

Like region 1, this region is also referred to as a *Salmonella* pathogenicity island (SPI). SPI-7, or the major Vi pathogenicity island, is ~134 kb in length and contains ~30 kb of integrated bacteriophage. Have a look at the CDSs within this region. As before notice any stable RNAs that may have acted as the phage integration site.

## Artemis Exercise 1 Part IV

Continuing on from the analysis of Region 3 or SPI-7 (the major Vi-antigen pathogenicity island) we are going to extract this region from the whole genome sequence and perform some more detailed analysis on it. We will aim to write and save new EMBL format files which will include just the annotations and DNA for this region.

[illegible]

The screenshot shows the Artemis genome browser interface. At the top, the menu bar includes File, Entries, Select, View, Goto, Edit, Create, Write, Run, Graph, and Display. Below the menu, the status bar indicates "Nothing selected". The main window displays a genomic track with various features. A yellow callout box points to the "Entry:" field, which has "no name" selected. A blue callout box points to the "misc\_feature" track, which shows a feature labeled "misc\_feature" with a range of 800 to 7200. A red callout box points to the "tRNA" track, which shows a feature labeled "tRNA" with a range of 800 to 7200. A green callout box points to the "Sequence" view, which shows the DNA sequence starting with "K A G R C T G I I V C P Y R T D C S G Q Q N N T D T S G # Q H P".

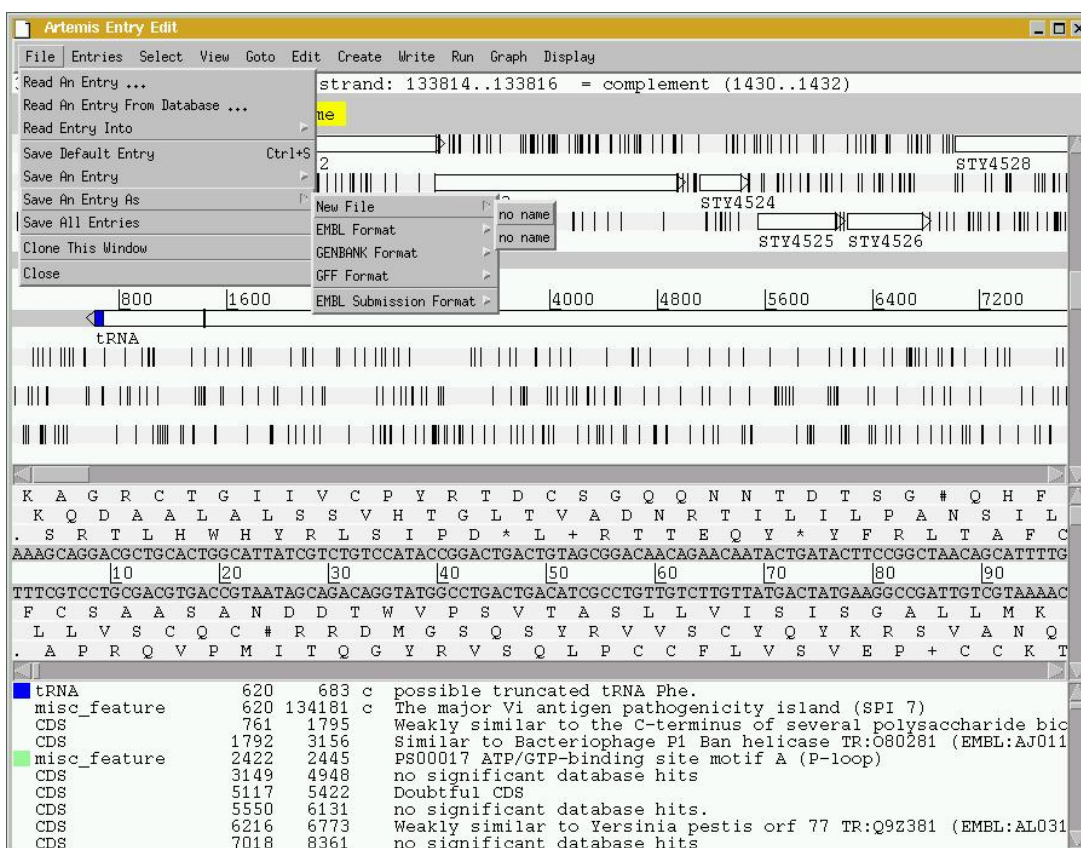
A new Artemis window will appear displaying only the region that you have highlighted

Note the bases have been renumbered from the first base you selected.



Note that the two entries on the grey Entry line are now denoted 'no name', they represent the same information in the same order as the original Artemis window but simply have no assigned name. Because the sub-sequence is now viewed in a new Artemis session, this prevents the original files from being over written (i.e. S\_typhi.dna and S\_typhi.tab). We will now save them as new files to avoid confusion. So click on the File menu then 'Save an entry as' and then 'New file'. Another menu will ask you to choose one of the entries listed. At this point they will both be called 'no name'. Left click on the top entry in the list. A window will appear asking you to give this file a name. Save this file as spi7.dna

Do the same again for the other unnamed entry and save it as spi7.tab



We are going to look at this region in more detail and to attempt to define the limits of the bacteriophage that lies within this region. Luckily for us all the phage-related genes within this region have been given a colour code number 12 (pink). We are going to use this information to select all the relevant phage genes using the Feature selector as shown below and then to define the limits of the bacteriophage.

First we need to create a new entry (click 'Create' then 'New Entry'). Another entry will appear on the entry line called, you guessed it, 'no name'. We will eventually copy all our phage-related genes into here.

1 Click 'Select' then 'Feature Selector'

2 Make sure the buttons are down

3 Set Key to 'CDS' and Qualifier to 'colour'

4 Type search term

5 Click to select features containing search term

6 Click to view selected features

7 Double click to bring feature into main view window

Artemis Entry Edit

File Entry Select View Goto Edit Create Write Run Graph Display

Nothing selected

Entry: All Ctrl+A

None Ctrl+N

By Key

CDS Features

Same Key

Open Read

Features i

Base Rang

Feature A

Toggle Se

Artemis Feature Selector

Select by:

Key: CDS

Qualifier: colour

Containing this text: 12

Ignore Case

Allow Partial Match

And by:

Amino acid motif:

Forward Strand Features

Reverse Strand Features

Select View Close

All features with key 'CDS' with qualifier 'colour' containing text '12'

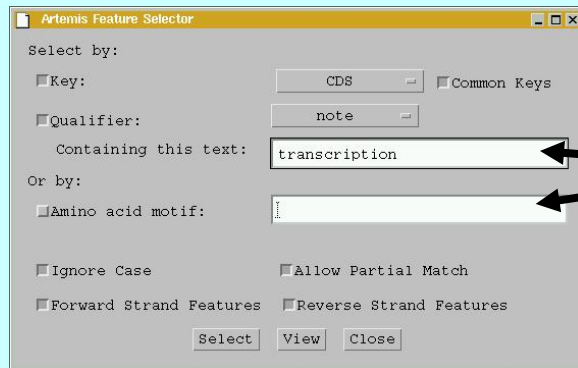
Feature	Start	End	Strand	Description
CDS	65082	65459	c	no significant database hits
CDS	65546	65764	c	Similar to Escherichia coli prophage P2
CDS	65832	66932	c	Similar to Bacteriophage P2 late gene
CDS	66929	67414	c	Similar to Bacteriophage P2 complete ge
CDS	67414	70194	c	Similar to Bacteriophage 186 protein G
CDS	70187	70306	c	Similar to Bacteriophage 186 Orf52 H TR
CDS	70321	70623	c	Similar to Bacteriophage P2 late gene
CDS	70678	71193	c	Similar to Bacteriophage P2 major tail
CDS	71203	72375	c	Similar to Bacteriophage P2 major tail
CDS	72910	73632	c	Similar to Salmonella typhimurium invasi
CDS	73830	74237	c	Similar to Bacteriophage P2 probable t
CDS	74244	75863	c	Similar to Bacteriophage P2 probable t
CDS	75860	76465	c	Similar to Bacteriophage P2 tail protei
CDS	76458	77366	c	Similar to Bacteriophage P2 baseplate
CDS	77353	77712	c	Similar to Bacteriophage P2 baseplate
CDS	77709	78287	c	Similar to Bacteriophage P2 baseplate
CDS	78356	78802	c	Similar to Bacteriophage P2 tail complet
CDS	78795	79226	c	Similar to Bacteriophage P2 tail complet
CDS	79322	79747	c	Similar to Bacteriophage P2 protein Lys
CDS	79747	80124	c	no significant database hits. Contains
CDS	80129	80593	c	Similar to Serratia marcescens putative
CDS	80619	80834	c	Similar to Serratia marcescens extracell

The genes listed in **6** are only those fitting your selection criterion. They can be copied or moved in to a new entry so we can view them in isolation from the rest of the information within spi7.tab.

Firstly in window **6** select all of the CDS shown by clicking on the 'select' menu and then selecting 'All'. All the features listed in window **6** should now be highlighted. To copy them to another entry (file) click 'Edit' then 'move selected Features To' then 'no name'. Close the two smaller feature selector windows and return to the SPI-7 Artemis window. You could rename the 'no name' entry as you did before. Temporarily remove the features contained in 'spi7.tab' file by left clicking on the entry button on the grey entry line. Only the phage genes should remain.

### Additional methods of selecting/extracting features using the Feature Selector

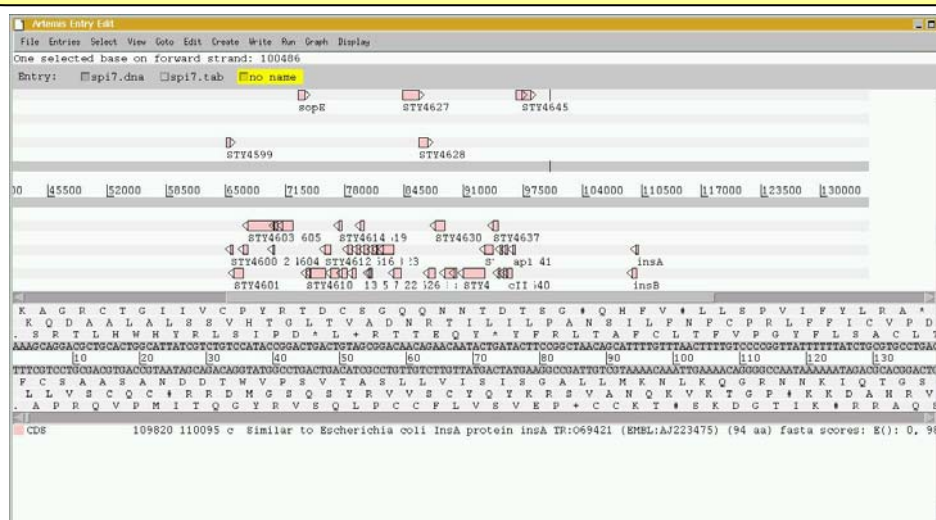
It is worth noting that the feature selector can be used in many other ways to select and extract subsets of features from the genome. If you have a closer look at the Feature selector you will also see that you can use search terms to select a class or all those features with a particular amino acid motif.



Space for a search term or amino acid motif

### Defining the extent of the prophage.

Even from this very cursory analysis it is clear from the selection that the prophage occupies a fairly discrete region within SPI-7 (see below). It is often useful to create a DNA feature to define the limits of this type of genome landmark. To do this use the left mouse button to click and drag over the region that you think defines the prophage. Click on the create menu and select 'Create feature from base range'. A feature edit window will appear. The default 'key' value given by Artemis when creating a new feature is 'CDS'. With this 'key' the newly created feature would automatically be put on the translation line. However, if we change this to 'misc\_feature' (an option in the key menu top left hand corner at the edit window) Artemis will place this feature on the DNA line. This is perhaps more appropriate and is easier to visualise. If you also add in a qualifier, such as '/label' and add text following the /label= ????, then click ok. That text will be used as a feature label to be displayed in the main sequence view panel.



To see how well you have done turn back on the spi7.tab and have a look at the genes located at either side of your selection. Go to and look at the CDS *samA*. In reality this gene was disrupted by the insertion of this bacteriophage. If you look at the FASTA results for this CDS you may be able to track the bases between which this phage inserted.

Your final task is to write out these files in EMBL format and create a merged annotation and sequence file in EMBL format:

1 Click 'File' then 'Save An Entry As'

2 EMBL Format

3 Select a file to save

The screenshot shows the Artemis Entry Edit window. The 'File' menu is open, and the 'Save An Entry As' option is highlighted. The window displays genomic data including coordinates, features, and a sequence view. The sequence view shows a DNA sequence with a highlighted region. The features table at the bottom lists various features such as tRNA, misc\_feature, and CDS, along with their coordinates and descriptions.

This will create two files one with the sequence and the other with the annotation in the directory within which you started Artemis. To create a complete EMBL file use the UNIX you covered earlier and 'cat' the files together



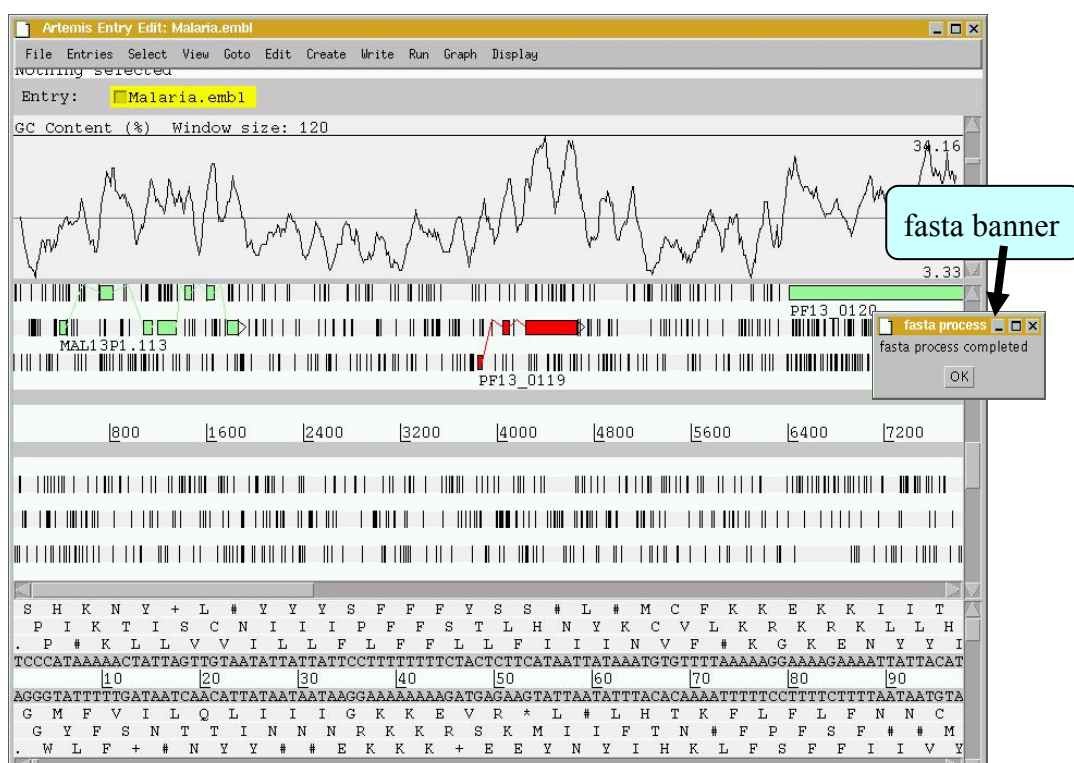
## Artemis Exercise 2

This exercise will look at a section of the Malaria genome. You will need to close down the last Artemis exercise if you haven't already done so. Then start a new Artemis Session, as before, using the file 'Malaria.embl' in the current directory (Module\_2\_Artemis). Unlike the Salmonella exercise, in this instance the annotation and sequence are contained within the same file 'Malaria.embl'

The sequence you are going to look at is a small region of contrived sequence (~21 kb) taken from *Plasmodium falciparum* chromosome 13. You will see 7 CDSs, some with multiple exons. As a gentle introduction to splicing we would like you to look at the genes named , PF13\_0119, MAL13P1.294 and PF13\_0061. They have only been partially characterised and may in fact be missing exons. Have a look at these CDSs and confirm, edit or dismiss the proposed gene models by using G+C content, database searches and looking for splice sites (**Appendix IX**).

G+C content is a very good indicator of coding capacity in Malaria. On average, the coding regions are ~23% G+C and the non-coding regions are ~19%. Have a look at the G+C content for this region by selecting the appropriate graph. Left click within the graph window and then select by clicking on the exons to see how this relates to the G+C peaks on the graph.

Note, we will cover the principals and methods of gene prediction in much more detail in a module 3.



To compare the three CDS with others currently in the public databases run a fasta search. Left click the CDS, click on the 'Run' menu and then 'Run fasta on selected features'. When the search is finished, a banner will appear saying 'fasta process completed' (see above). The search may take a couple of minutes to run.

To view the search results click 'View' then 'Search Results' then 'fasta results'. The results will appear in a scrollable window. You could also view these results in your Netscape Browser window as in the previous exercise.

How does your predicted gene model for this CDS compare with proteins pulled out of the public databases? Is it possible that there are additional exons not featured in the current model.

If you think that there are additional exons that should have been included in the gene model you should add them to it. Using GC content and results from your database search as guides roughly draw in where you think the additional exon(s) lie:

To create additional exons:

Select the region you think represents the exon by holding down the left mouse button and dragging the cursor over the region of interest. Then click the 'Create' menu and select 'Create feature from base range'. A new blue CDS feature will appear on the appropriate frame line (See below).

The screenshot shows the Artemis Entry Edit window for 'Malania.embl'. The window displays a genomic map with a GC content plot at the top, a gene model below it, and a sequence view at the bottom. A new CDS feature is being added to the gene model. The sequence view shows the following sequence:

```

C S * G N # C R A S F S S I K W L V H I Y T M R
L L R Q L M + S + L F I N E # M V S T H I Y N E K C # V F I Y L F
G C T C T G A G G C A A C T A A T G T A G A G C T A G C T T T T C A T C A A T A T A A T G G T T A G T A C A C A T A T A C A A T G A G A A A T G T T A A G T A T T T A T T A T T A
0 4620 4630 4640 4650 4660 4670 4680 4700
C G A G A A C T C G T T G A T T A C A T C G A T C G A A A G T A G T T A A T T A C C A A T C A T G T A T A T A T G T T A C T C T T A C A A T T C A T A A A T A A A T A A A T
A R S A V L T S S A K * * N N F P # Y V V I C H S I N L I # K N L
S K L C S I Y L + S K M L # I T L V C I Y L S F H # T N I # K
E Q P L + H L A L K E D I I L H N T C M Y V I L F T L Y K N I #

```

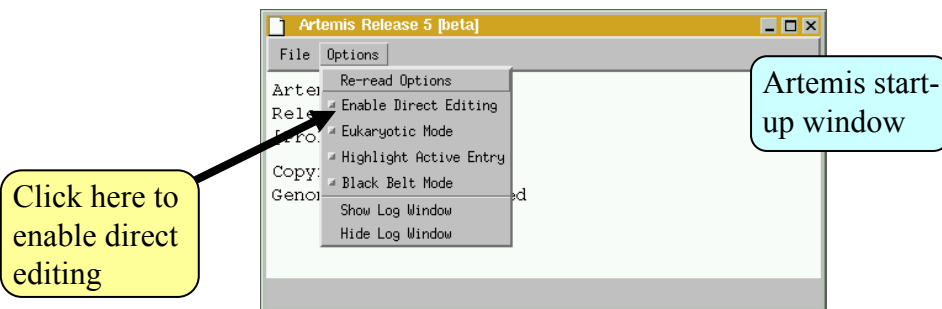
Annotations in the image:

- 1**: Select both the original gene-model and the new CDS feature, which is to be merged with it to form a new exon.
- 2**: Click Edit.
- 3**: Merge Features.

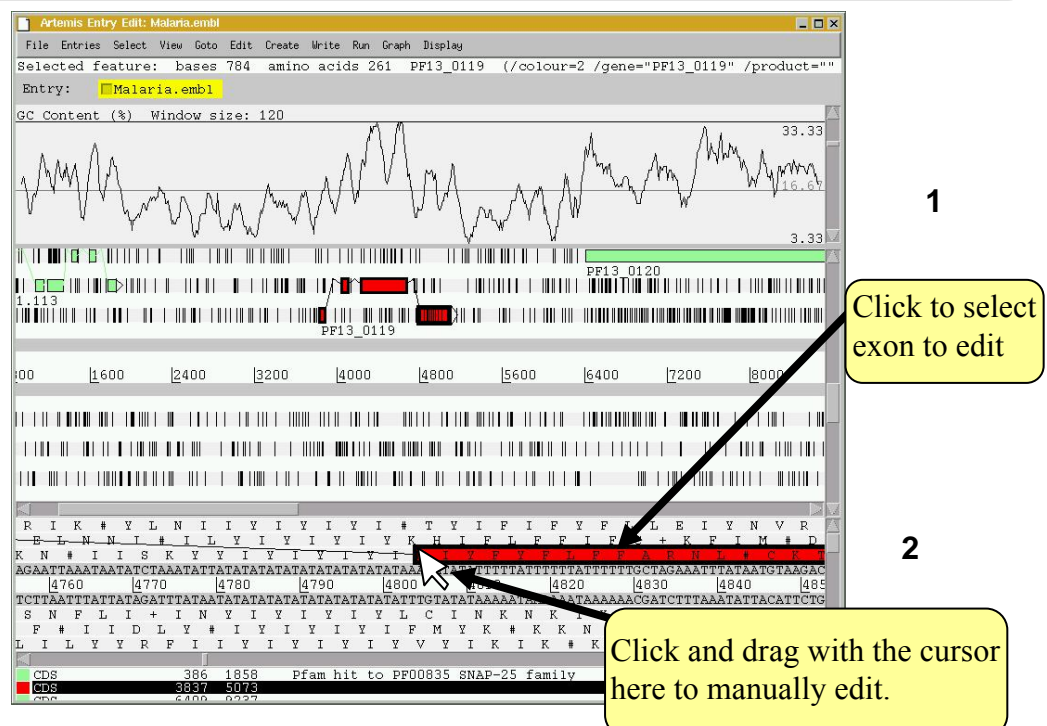
Tip, to select more than one feature (of any type) you must hold the shift key down.

The new CDS feature can then be merged with the original gene model as shown above.

A small window will appear asking you whether you are sure you want to merge these features. Another window will then ask you if you want to 'delete old features'. If you click 'yes' the CDS features you have just merged will disappear leaving the single merged CDS. If you select 'no' all of the three CDS features (the two CDSs that you started with plus the merged feature) will be retained.



You may notice after you performed the merge function that one of the exons has subsequently jumped into another reading frame. Artemis automatically splices the CDS and so if the exon boundaries have an additional partial codon then any following exon will be pushed into another reading frame to account for this. To correct this you can edit the exon boundaries directly by turning on manual editing in the options menu of the Artemis start-up window, (as shown above). This will now allow you to edit the start and end positions of the feature boxes by using the left mouse button. Click and hold down the cursor over the first or last base of any feature and then drag the mouse. The feature box should move as you drag it (see below). This can be a little tricky so please ask



When manually editing your exons you can should look out for appropriate splice donor and acceptor sites. See below for a small list and **Appendix IX** for details of known acceptor and donor motifs for Malaria splice sites.

Once you are happy with your newly created exon re-run the fasta search and see how this compares with the other hits in the public databases. If there are more exons to mark up try and complete the gene model.

The three example CDS to analyse were selected because they have very good database hits. This obviously makes the task of making the gene model far easier. However, several of the other CDS in this region have no significant database hits. If you have time you may want to have a look at these too.



# Module 2

# Comparative Genomics

## Introduction

The Artemis Comparison Tool (ACT), also written by Kim Rutherford, was designed to extract the additional information that can only be gained by comparing the growing number of genomes from closely related organisms.

ACT is based on Artemis, and so you will already be familiar with many of its core functions.

ACT, is essentially composed of three layers or windows. The top and bottom layers are mini Artemis windows (with their inherited functionality), showing the linear representations of the genomes with their associated features. The middle window shows red blocks, which span this middle layer and link conserved regions within the two genomes, above and below.

Consequently, if you were comparing two identical genome sequences you would see a solid red block extending over the length of the two sequences in this middle layer. If insertions were present in either of the genomes, they would show up as breaks between the solid red conserved regions. Data used to draw these red blocks and link conserved regions is generated by running pairwise BlastN or tBlastX comparisons of the genomes (details of how this is done are outlined in Appendix II and can be obtained from the ACT user manual:

[http://www.sanger.ac.uk/Software/ACT /manual/](http://www.sanger.ac.uk/Software/ACT/manual/)).

## Aims

The aim of this Module is for you to become familiar with the basic functioning of ACT by using a series of worked examples. Some of these examples will touch on exercises that were used in previous Modules, this is intentional. Hopefully, as well as introducing you to the basics of ACT this Module will also show you how ACT can be used for not only looking at genome evolution but also to backup, or question, gene models and so on.

# 1. Starting up the ACT software

Make sure you're in the correct directory **Comparative Genomics Module 5**.

Then type

**act & [return]**

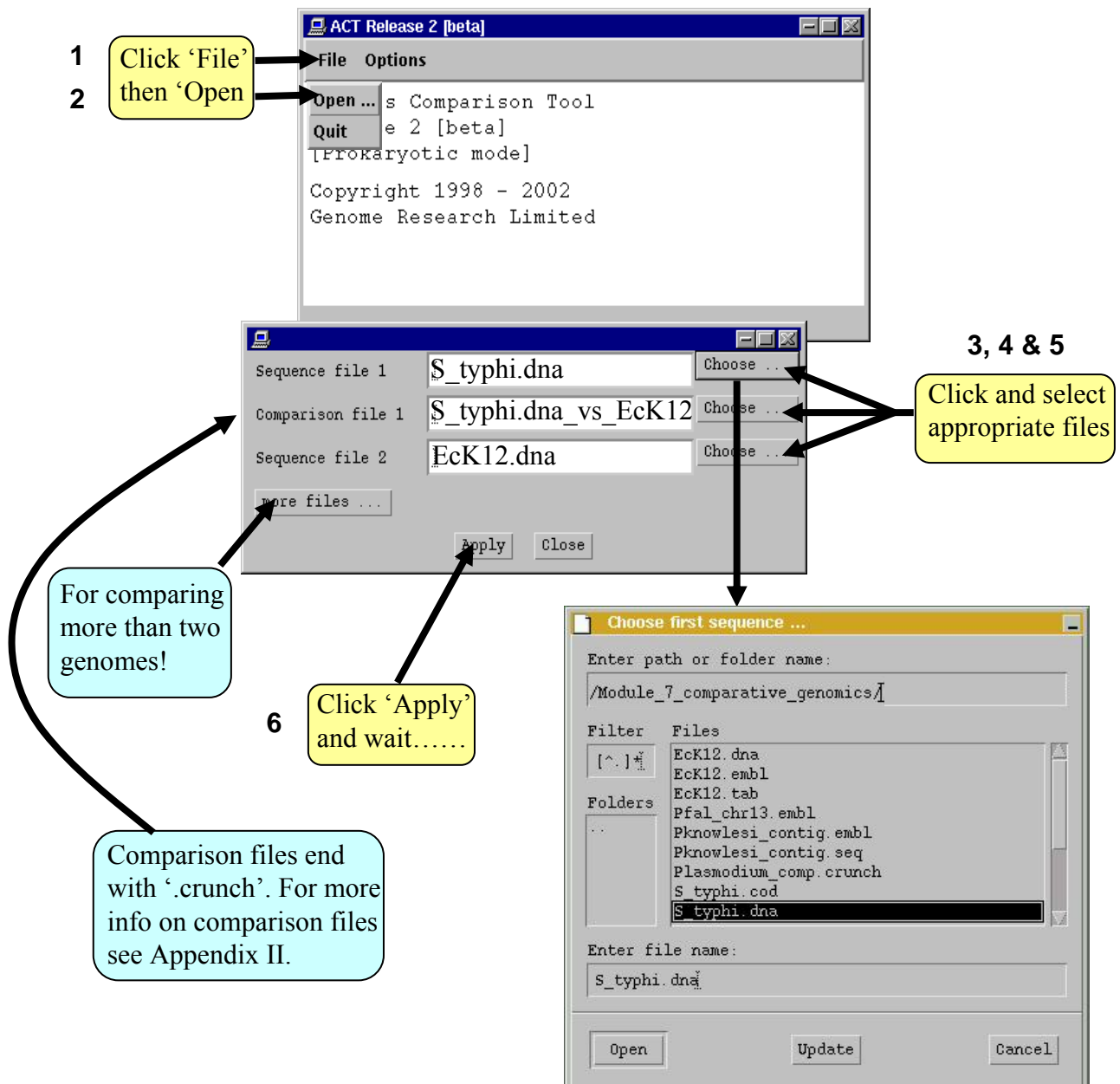
A small start up window will appear.

Now let's load up a *S. typhi* versus *Escherichia coli* comparison.

The files you will need for this exercise are: *S\_typhi.dna*

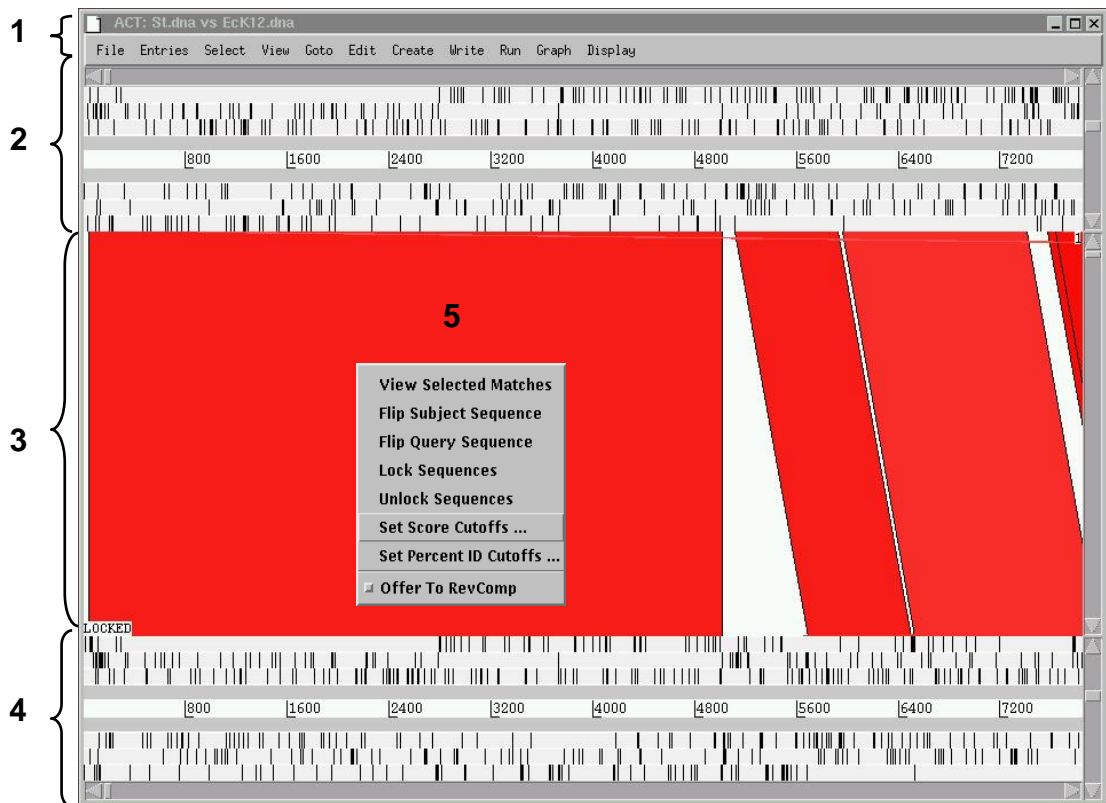
*S\_typhi.dna\_vs\_EcK12.dna.crunch*

*EcK12.dna*

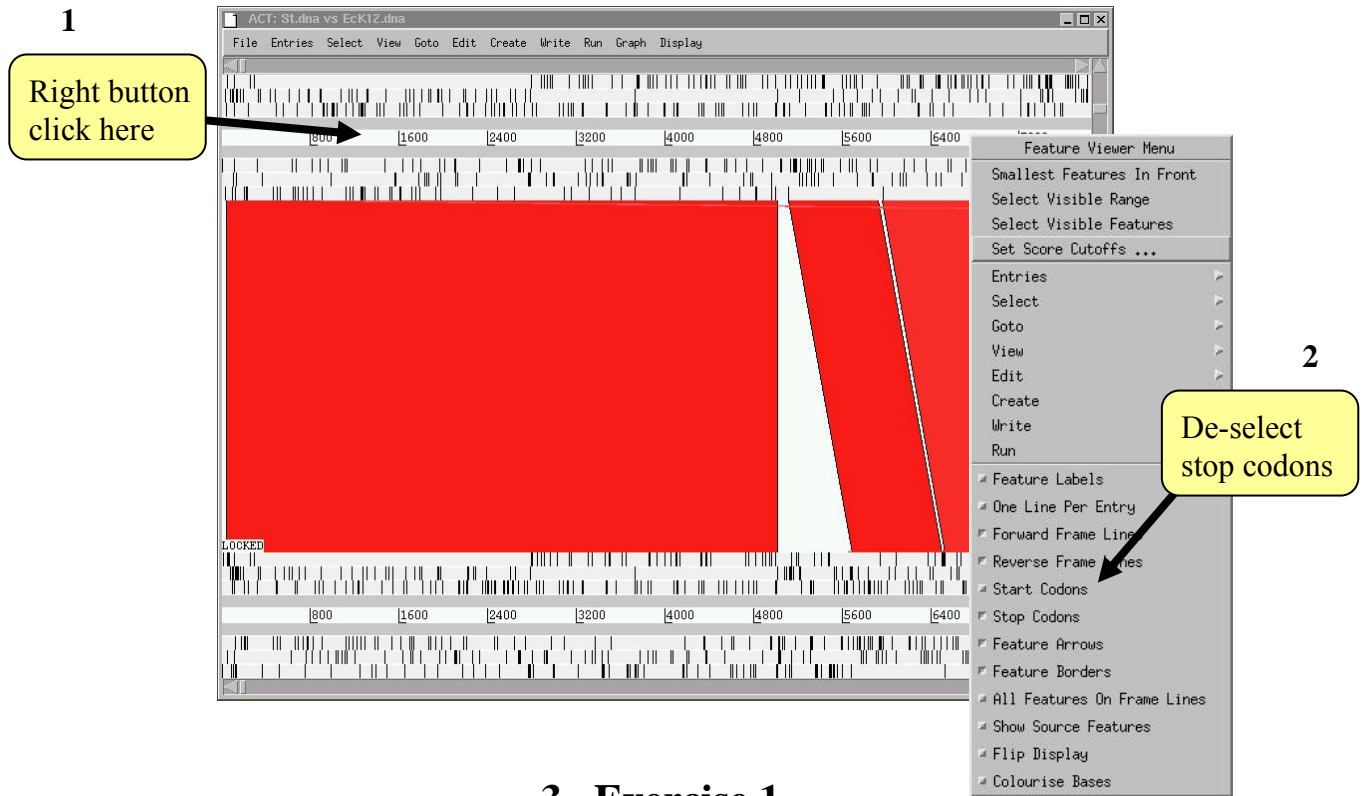


## 2. The basics of ACT

You should now have a window like this so let's see what's there.



1. Drop-down menus. These are mostly the same as in Artemis. The major difference you'll find is that after clicking on a menu header you will then need to select a DNA sequence before going to the full drop-down menu.
2. This is the Sequence view panel for 'Sequence file 1' (Subject Sequence) you selected earlier. It's a slightly compressed version of the Artemis main view panel. The panel retains the sliders for scrolling along the genome and for zooming in and out.
3. The Comparison View. This panel displays the regions of similarity between two sequences. Red blocks link similar regions of DNA with the intensity of red colour directly proportional to the level of similarity. Double clicking on a red block will centralise it.
4. Artemis-style Sequence View panel for 'Sequence file 2' (Query Sequence).
5. Right button click in the Comparison View panel brings up this important ACT-specific menu which we will use later.



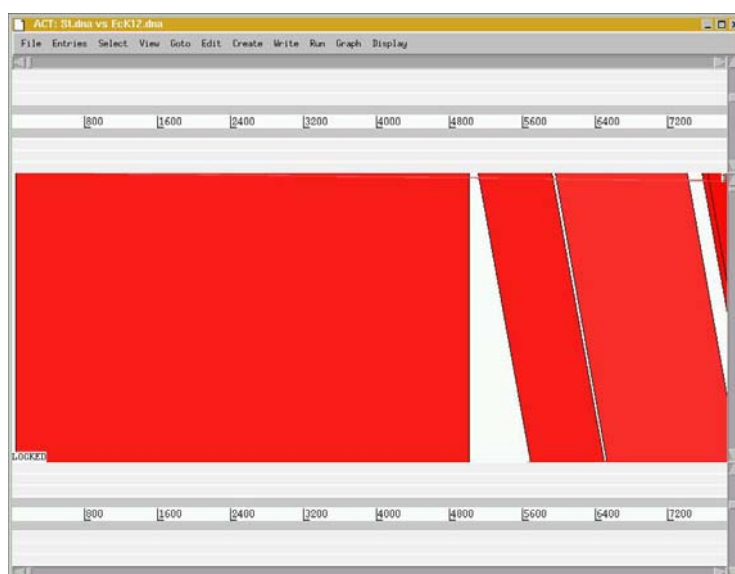
### 3. Exercise 1

#### Introduction & Aims

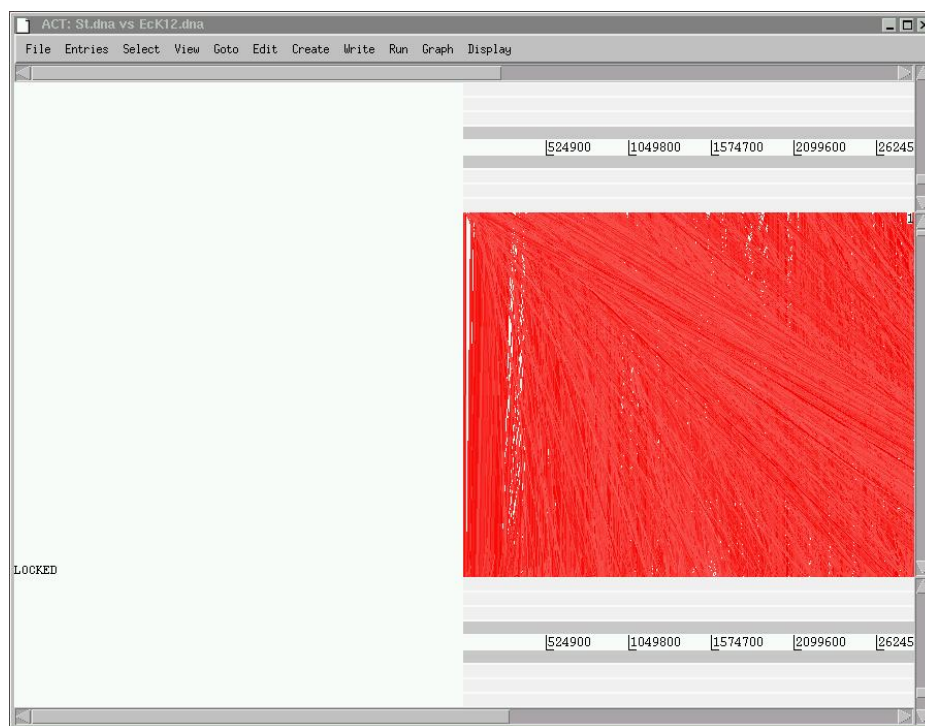
In this first exercise we are going to explore the basic features of ACT. Using the ACT session you have just opened we firstly are going to zoom outwards until we can see the entire *S. typhi* genome compared against the entire *E. coli* K12 genome. As for the Artemis exercises we should turn off the stop codons to clear the view and speed up the process of zooming out.

The only difference between ACT and Artemis when applying changes to the sequence views is that in ACT you must click the right mouse button over the specific sequence that you wish to change, as shown above.

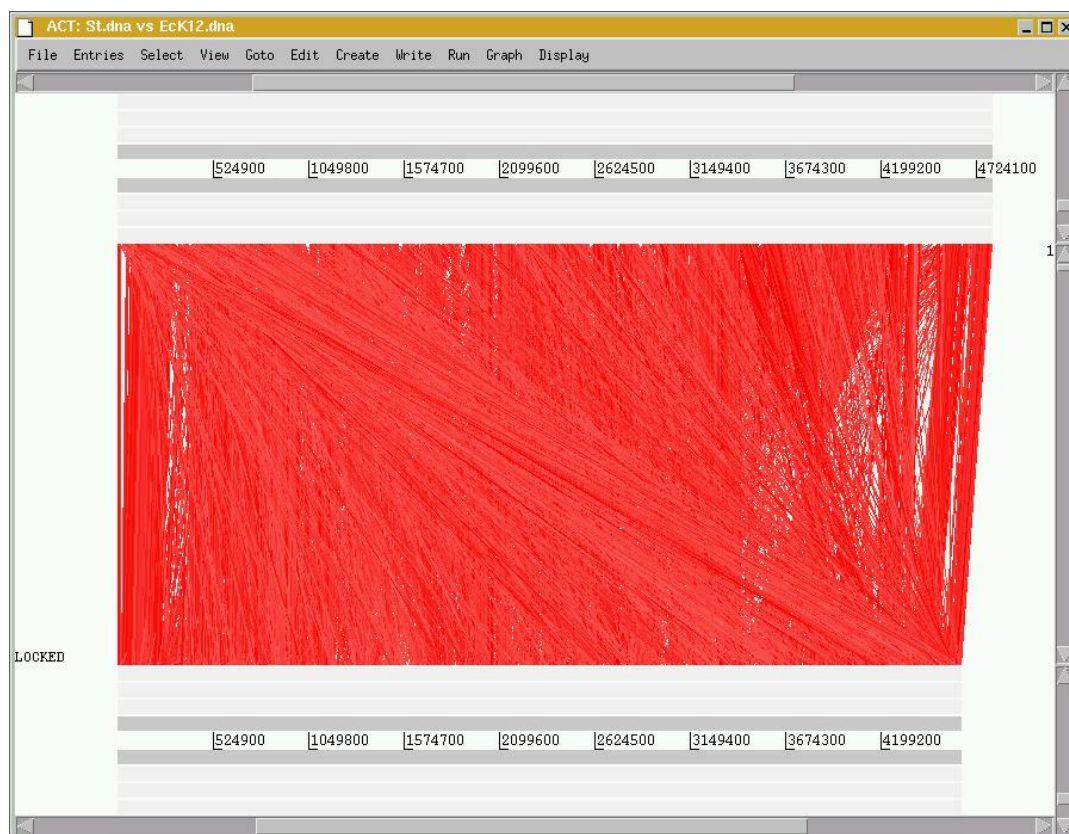
Now turn the stop codons off in the other sequence too. Your ACT window should look something like the one below:



Use the vertical sliders to zoom out. Drag or click the slider downwards from one of the genomes. The other genome will stay in synch.

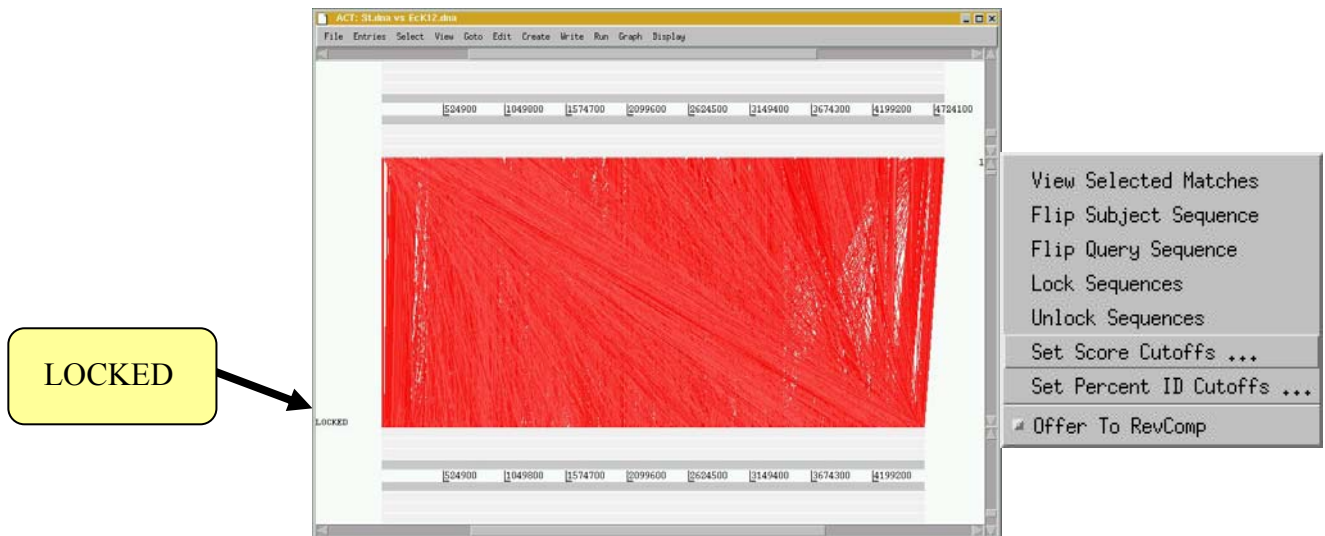


Once zoomed out your ACT window should look similar to the one shown above. If the genomes in view fall out of view to the right of the screen, use the horizontal sliders to scroll the image and bring the whole sequence into view, as shown below. You may have to play around with the level of zoom to get the whole genomes shown in the same screen as shown below.

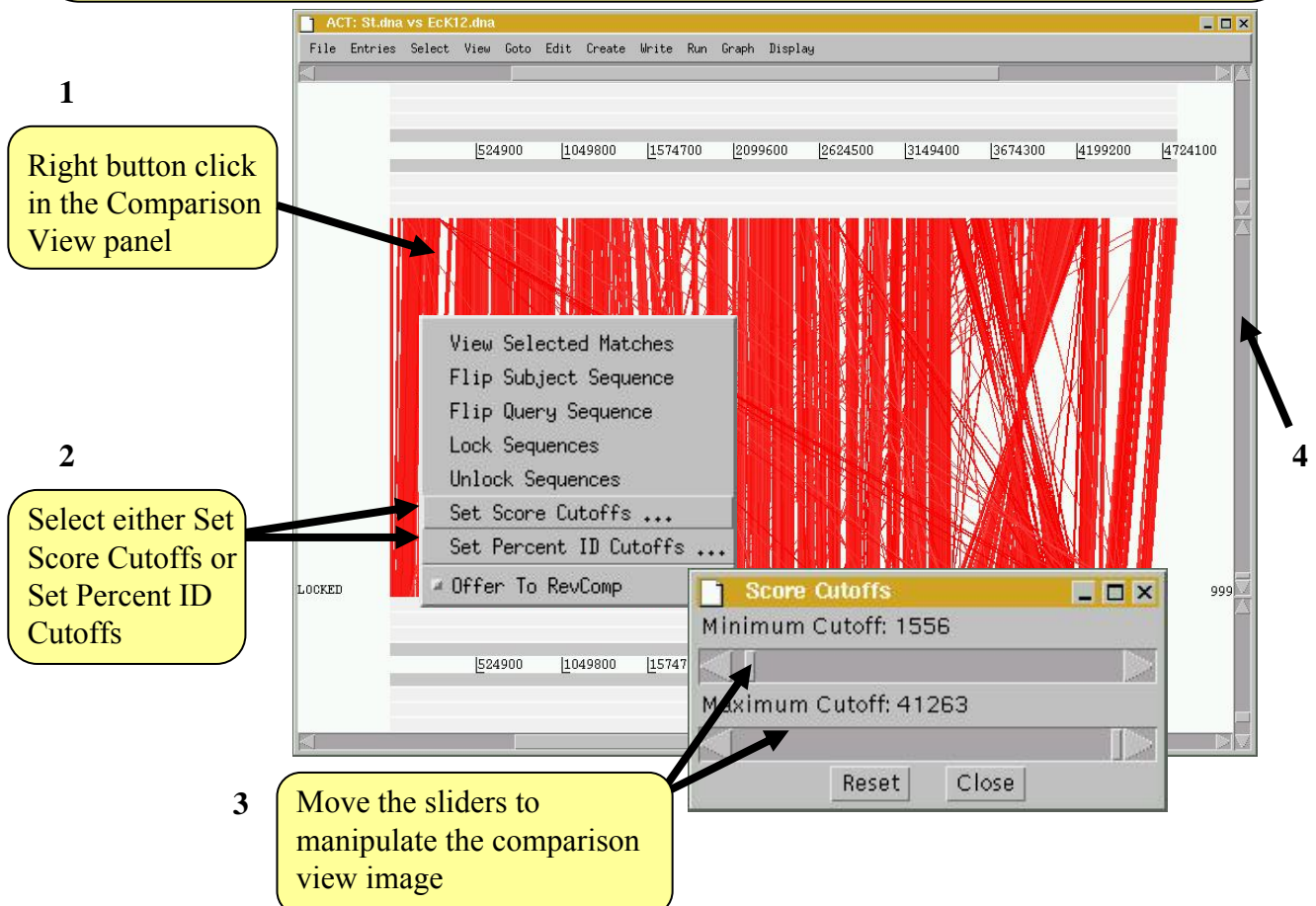


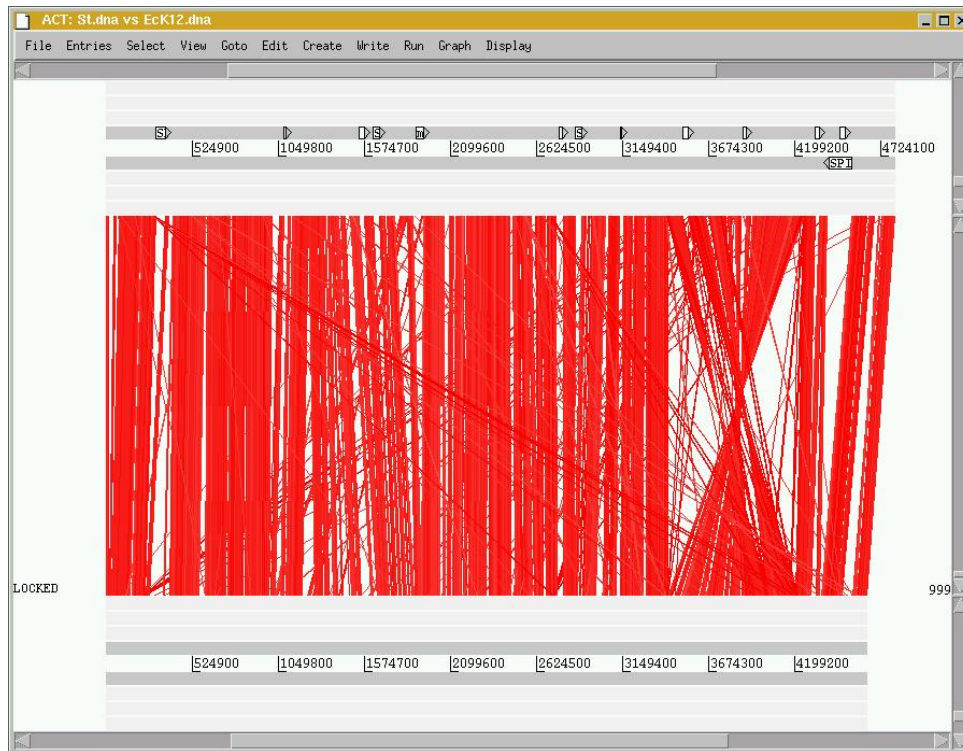


Notice that when you scroll along with either slide both genomes move together. This is because they are 'locked' together. Right click over the middle comparison view panel. A small menu will appear, select Unlock sequences and then scroll one of the horizontal sliders. Notice that 'LOCKED' has disappeared from the comparison view panel and the genomes will now move independently



You can optimise your image by either removing 'low scoring' (or percentage ID) hits from view, as shown below **1-3** or by using the slider on the the comparison view panel (**4**). The slider allows you to filter the regions of similarity based on the length of sequence over which the similarity occurs, sometimes described as the "footprint".





#### 4. Things to try out in ACT

Load into the top sequence (*S.typhi*) a '.tab' file called 'laterally.tab'. You will need to use the 'File' menu and select the correct genome sequence ('*S.typhi.dna*') before you can read in an entry. If you are zoomed out and looking at the whole of both genomes you should see the above. The small white boxes are the regions of atypical DNA covering regions that we looked at in the first Artemis exercise. It is apparent that there is a backbone sequence shared with *E. coli* K12. Into this various chunks of DNA, specific the *S. typhi* (with respect to *E. coli* K12) have been inserted.

#### 5. More things to try out in ACT

1. Double click red boxes to centralise them.
2. Zoom right in to view the base pairs and amino acids of each sequence.
3. Load annotation files into the sequence view panels.
4. You could load in the appropriate '.tab' files for each genome (*S\_typhi.tab* and *EcK12.tab*) and view the annotation of a particular region. Also try using some of the other Artemis features eg. graphs etc.
5. Find an inversion in one genome relative to the other then flip one of the sequences.

Once you have finished this exercise remember to close this ACT session down completely before starting the next exercise

## 6. Exercise 2 Part I:

### *Plasmoidum falciparum* and *Plasmodium knowlesi*: Genome Comparison

#### Introduction

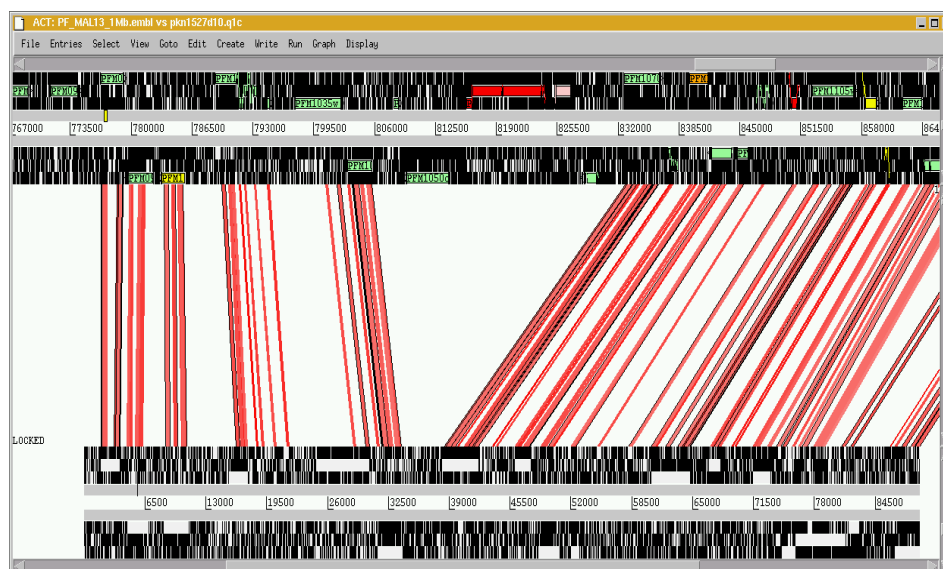
The parasite *P. falciparum* is responsible for hundreds of millions of cases of malaria and causes over 1 million deaths every year. Treatment and control have become difficult with the spread of drug-resistant malaria strains across the endemic countries in the world and there has been a major emphasis on research as part of our search for new drugs / vaccine candidates to fight against malaria. The analysis of the whole genome of *P. falciparum* has been completed and is made publicly available by the Malaria Genome Sequencing Consortium . Several animal models of malaria have also been used by researchers to study several aspects of malaria biology / host-parasite interactions. Sequences representing partial genomes of some of these model malaria parasites are also available now. This allows us to perform comparative analysis of the genomes of malaria parasites and understand the basic biology of their parasitism, based on the similarities / dissimilarities between the parasites at DNA / predicted protein level.

#### Aim

You will be looking at the comparison between a genomic DNA fragment of the primate malaria *P. knowlesi* and the previously annotated chromosome 13 of *P. falciparum* . By comparing the two genomic fragments you will be able to study the degree of conservation of gene order and identify new genes in *P. knowlesi* genome. As part of the exercise you will also identify any gross dissimilarity visible between the the two genomic fragments and finally, predict/ modify the gene model for one multi-exon gene in *P. knowlesi* genomic fragment.

The files that you are going to need are:

Pfal_chr13.embl	- annotation file with sequence
Pknowlesi_contig.seq	- sequence file (without annotation)
Pknowlesi_contig.embl	- annotation file with sequence
Plasmodium_comp.crunch	- tblastx comparison file



*P. falciparum*  
chr 13 (fragment)

*P. knowlesi*  
contig

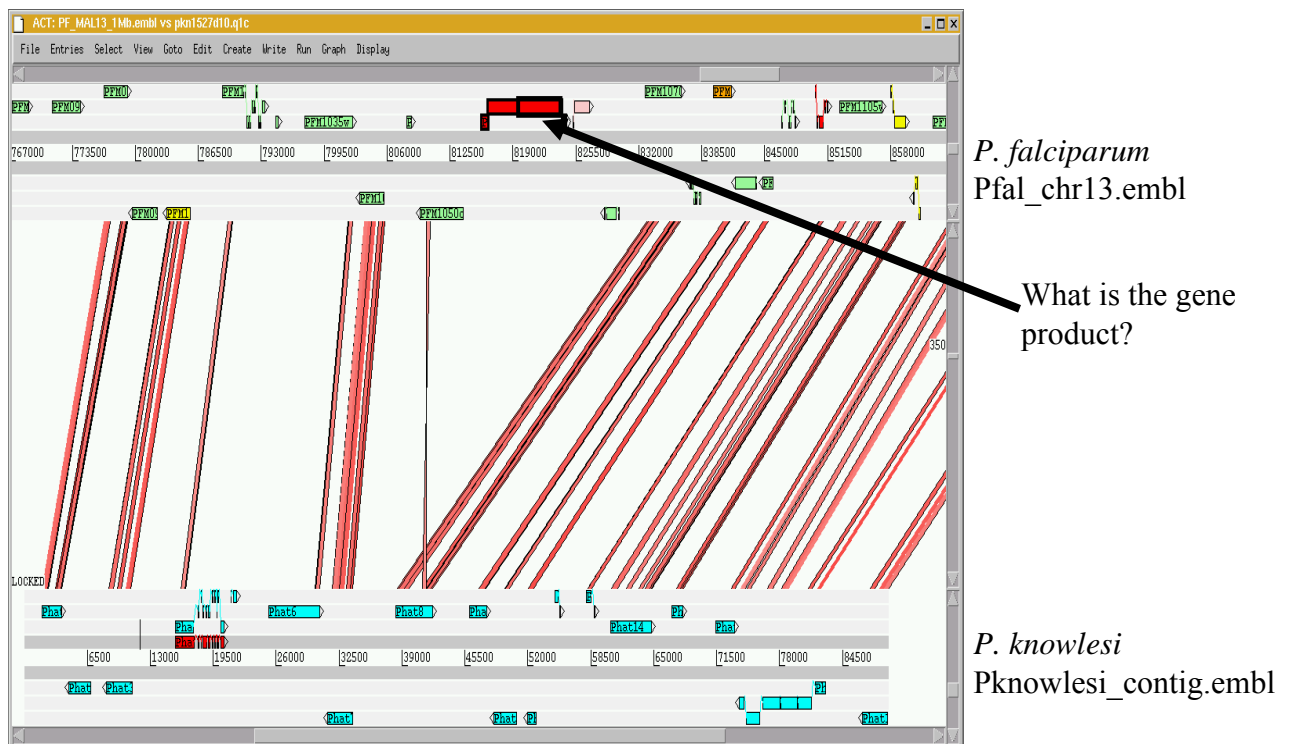
Comparison of *P. knowlesi* contig and the annotated chromosome 13 fragment of *P. falciparum*



## Exercise 2 Part II

### Conservation of gene order (synteny)

- In the ACT start up window load up the files Pfal\_chr13.embl, Pknowlesi\_contig.seq and the comparison file Plasmodium\_comp.crunch
- Use the slider on either sequence view panel to obtain a global view of the genome comparison. Also used the slider on the comparison view panel to remove the 'shorter' similarity hits. What effects does this have?
- Can you see conserved gene order between the 2 species?
- Can you see any region where similarity is broken up? Zoom in and look at some of the genes encoded within this unique region in file: Pfal\_chr13.embl (top sequence)
- Example location: **Pfal\_chr13.embl**, 815823..829969
- What are the predicted products of the genes assigned to this unique location? View the details by clicking on the feature, and then select 'Edit selected feature' from the 'Edit' menu after selecting the appropriate CDS feature.
- Can you identify a few putative genes in *P. knowlesi* contig, based on their conserved and syntenic nature with *P. falciparum* chromosome 13? Activate / inactivate stop / start codons in an entry, using the right click button on the mouse. This will allow you to see any potential ORFS.
- Any thoughts about the possible biological relevance of the comparison?



## Exercise 2 Part III

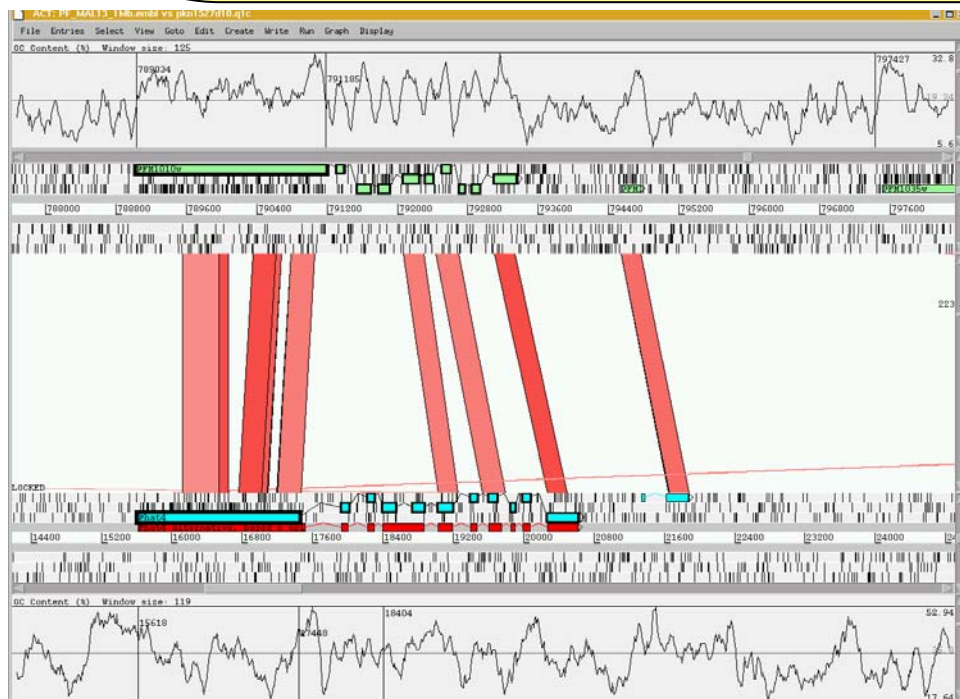
### Prediction of gene models:

There are several computer algorithms covered earlier in Module 3 that predict gene models, based on training the algorithm with previously known gene sets with previously known experimentally verified exon-intron structures (in eukaryotes). However, no single programme can predict the gene structure with 100% accuracy and one needs to curate / refine the gene models, generated by automated predictions. We have generated automated gene models for the *P. knowlesi* contig, using PHAT (Pretty Handy Annotation Tool, a gene finding algorithm, see in Mol. Biochem. Parasitol. 2001 Dec;118(2):167-74) and the automated annotation is saved in Pknowlesi\_contig.embl.

- Zoom into the *P. falciparum* gene labelled PFM1010w shown below. Can you compare the 2 gene models and identify the conserved exon(s) between the 2 species?
- Use the slider on the comparison view panel to include some 'shorter' similarity hits. Can you now identify all the conserved exons of the PFM1010w orthologue in the *P. knowlesi* contig? (For the time being, disregard the misc\_feature for 'Phat4', coloured in red in the 'Pknowlesi\_contig.embl' file )
- Open the 'GC Content (%)' window from 'graph' menu for both the entries. Can you relate the exon-intron boundaries to GC-content for the *P. falciparum* gene labelled PFM1010w? Is it also applicable to the gene model 'Phat4' in the *P. knowlesi* contig?
- Example regions:

**Pfal\_chr13.embl, 789034..793351**

**Pknowlesi\_contig.embl, 15618..20618**



*P. falciparum*  
Pfal\_chr13.embl

*P. knowlesi*  
Pknowlesi\_contig.embl

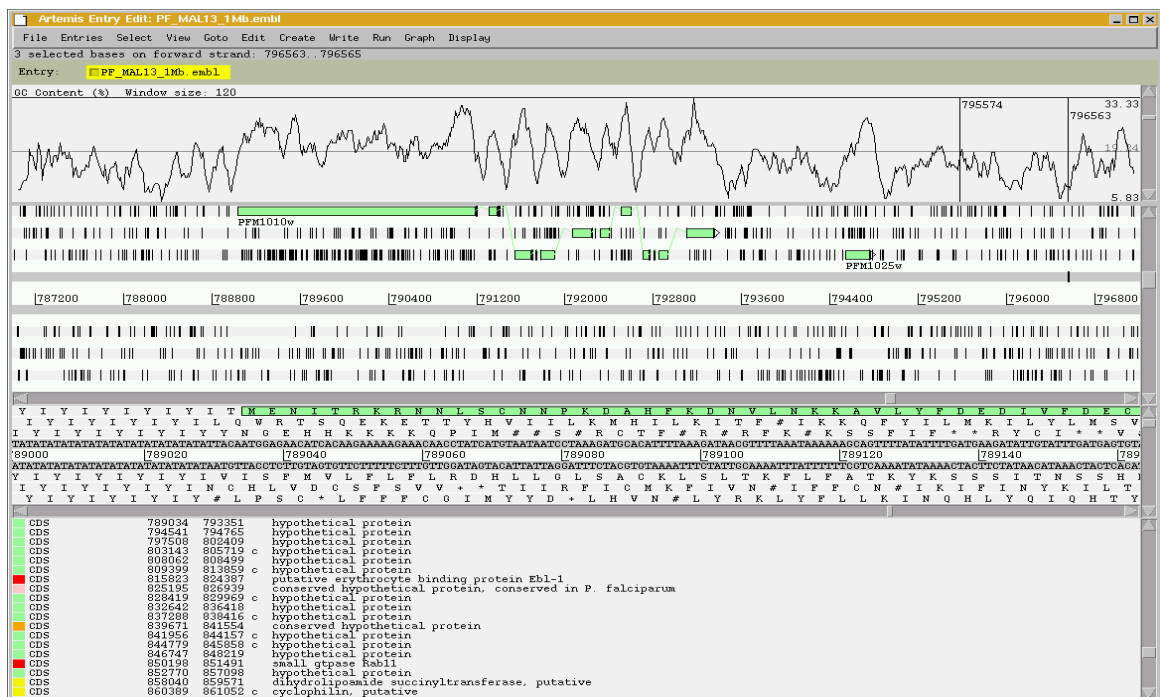
Comparison between orthologous genes in *P. falciparum* and *P. knowlesi*

## Exercise 2 Part IV

### Gene models for multi-exon genes in *P. falciparum*:

- Use 'File' menu to select entry 'Pfal\_chr13.embl' and select 'Edit In Artemis' to bring up an Artemis window.
- In Artemis window, use 'Graph' menu and switch 'on' the 'GC Content (%)' window.
- Use 'Goto' menu to select 'Navigator' window and within the Navigator window, select 'Goto Feature With This qualifier value' and type 'PFM1010w', click then close the dialogue box.
- Go through the annotated gene model for 'PFM1010w' and have a look at the the exon-intron boundaries and compare with the splice site sequences from *P. falciparum* given in Appendix IX.
- Also have a glance through a few other gene models for multi-exon genes and have a look at the intron sequences as well. Can you find any common pattern in the putative intron sequences? Hint – look at the complexity of the sequence
- You can delete exon(s) of any gene by selecting the exon(s) and then choosing 'Delete Selected Exons' from 'Edit' menu. Similarly, you can add an exon to a particular gene by co-selecting the exon and the gene (CDS features) followed by selecting 'Merge Selected Features' from the 'Edit' menu.
- Example regions:

**Pfal\_chr13.embl, 789034..793351, 657638..660023, 672361..673753**



Example location: 789034..793351, in Pfal\_chr13.embl

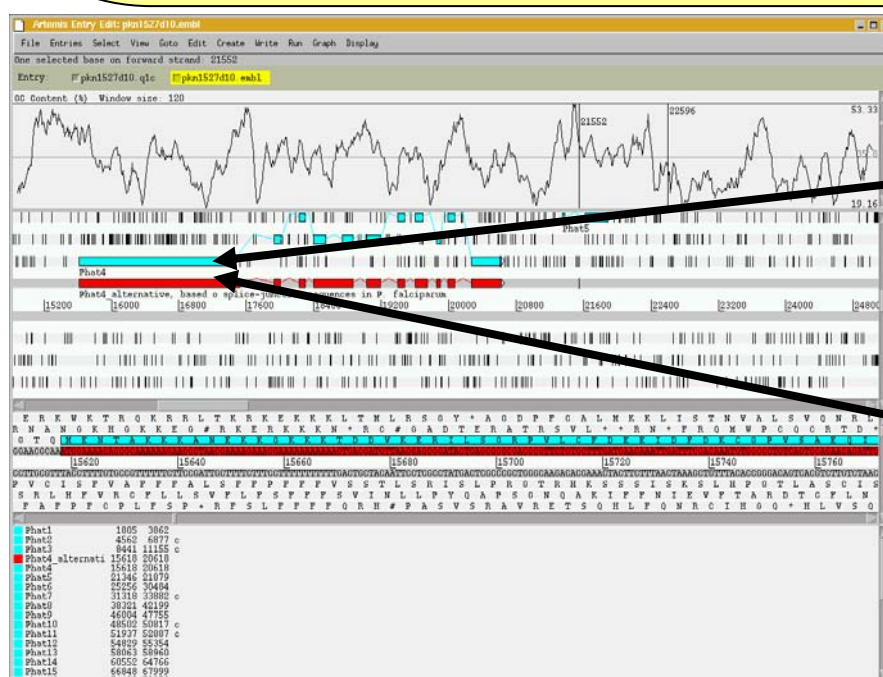
## Exercise 2 Part V

### Curation of gene models in *P. knowlesi*:

We are now going to edit the gene model for *P. knowlesi*.

- Use 'File' menu from the ACT displaying *P. falciparum* and *P. knowlesi* to select entry 'Pknowlesi\_contig.embl' and select 'Edit In Artemis' to bring up an Artemis window.
- Within the Artemis window, use 'Graph' menu and switch 'on' the 'GC Content (%)' window.
- Use 'Goto' menu to select 'Navigator' window and within the Navigator window, select 'Goto Feature With This Text' and type 'Phat4'.
- Go to the first ACT window, and use the 'Options' menu to select 'Enable Direct Editing'.
- Go through the gene model of 'Phat4' and have a glance through the exon-intron boundaries. Can you suggest any alternative gene model, after consulting the Table provided in Appendix IX, containing several examples of experimentally verified splice site sequences for *P. falciparum*?
- Example modifications:

Have a look at the 'misc\_feature', coloured in red (location: 15618..20618). Can you spot any difference in the red gene model of 'Phat4' at the exon-intron boundaries? Select the red feature, click on 'Edit' menu and select 'Edit Selected Features' and in the new window that pops out, change the 'Key' from misc-feature to 'CDS' and click on 'OK' button to close the window. Now you can compare the automatically created blue gene model and the curated red gene models at protein level and predict any alternative splicing pattern.



Automated gene prediction for hypothetical gene 'phat4'

Can you curate the 'Phat4' gene model and suggest any alternative splicing pattern such as the red model?

Example location: 15618..20618, in Pknowlesi\_contig.embl

## Exercise 3

### Introduction

Having familiarised yourselves with the basics of ACT, we are now going to use it to look at a region of synteny between *T. brucei* and *Leishmania*.

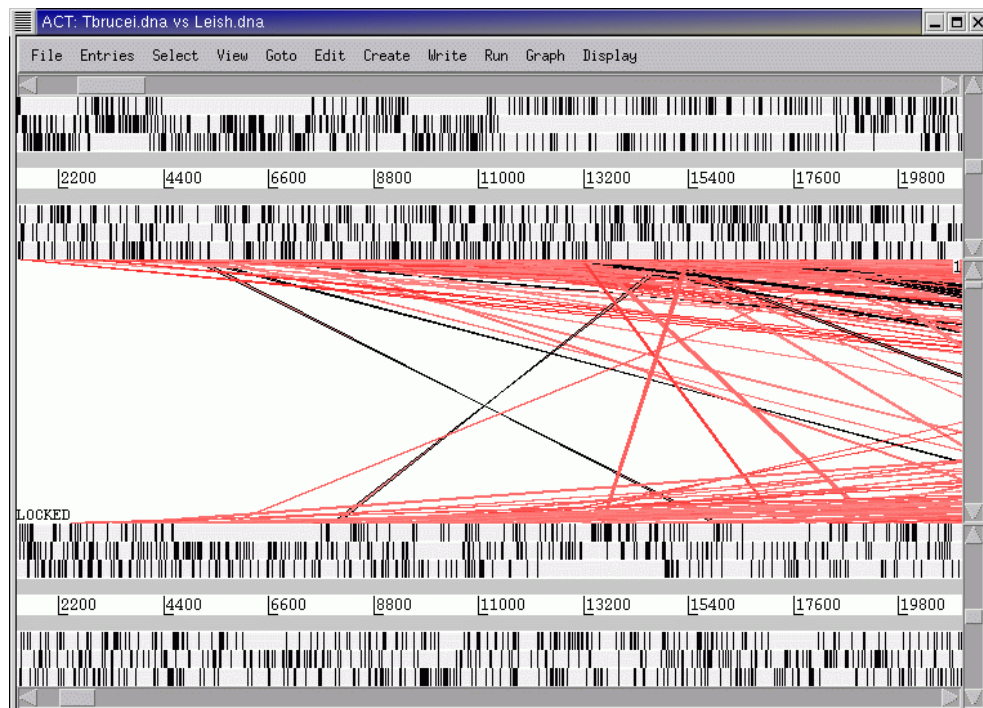
### Aim

By looking at a comparison of the annotated sequences of *T. brucei* and *L. major* you will be able to analyse, in detail, those genes that are found in both organisms as well as spot the differences. You will also see how act can be used to study the different chromosome architecture of these two parasite species.

The files that you are going to need are:

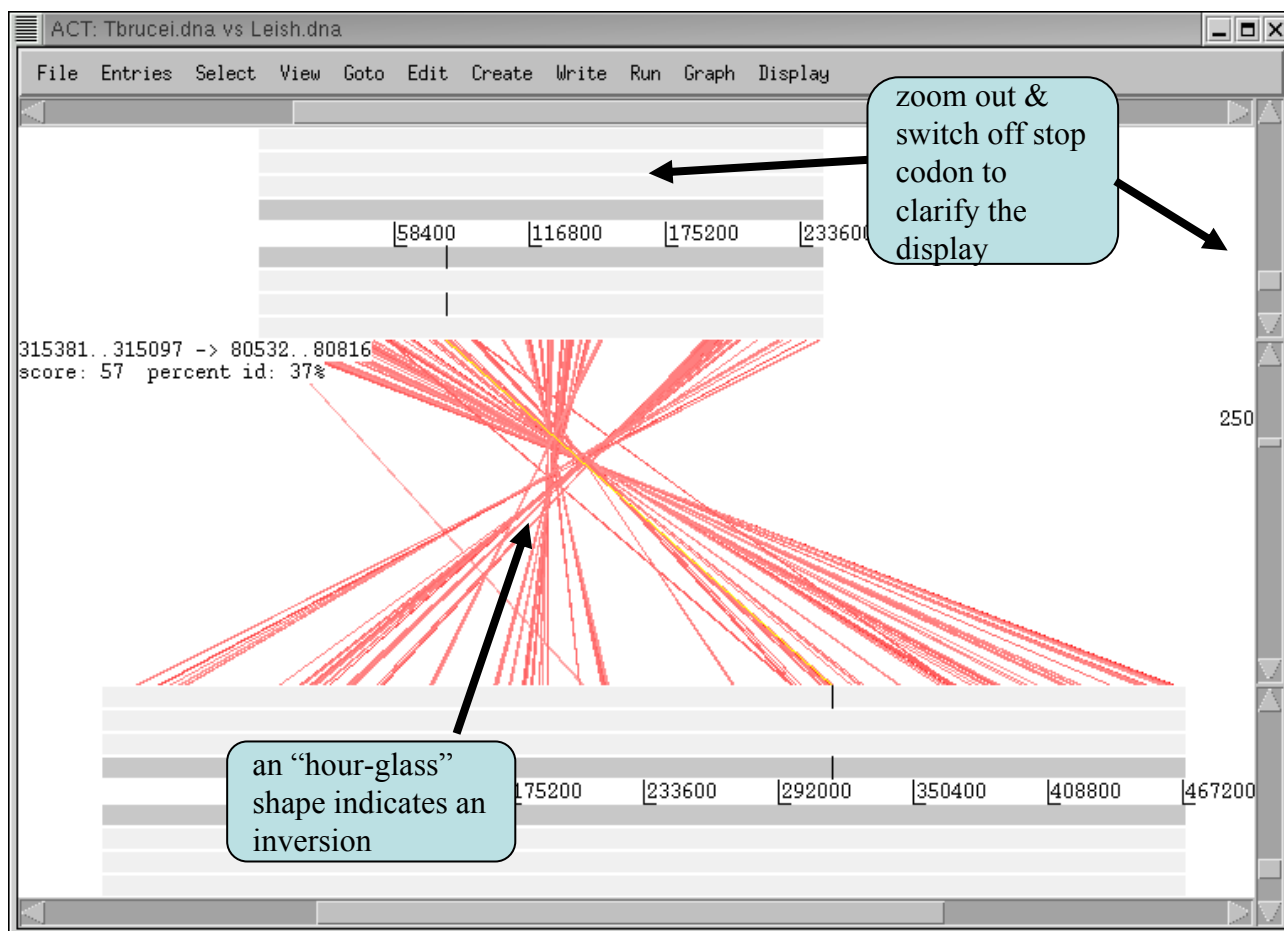
Tbrucei.dna	- <i>T. brucei</i> sequence
Tbrucei.embl	- <i>T. brucei</i> annotation
Leish_vs_Tbrucei.tblastx	- comparison file
Leish.dna	- <i>L. major</i> sequence
Leish.embl	- <i>L. major</i> annotation

First, load up the sequence files for *T. brucei* and *L. major* and the comparison file in ACT.





Next, you need to find the regions of synteny between the sequences.



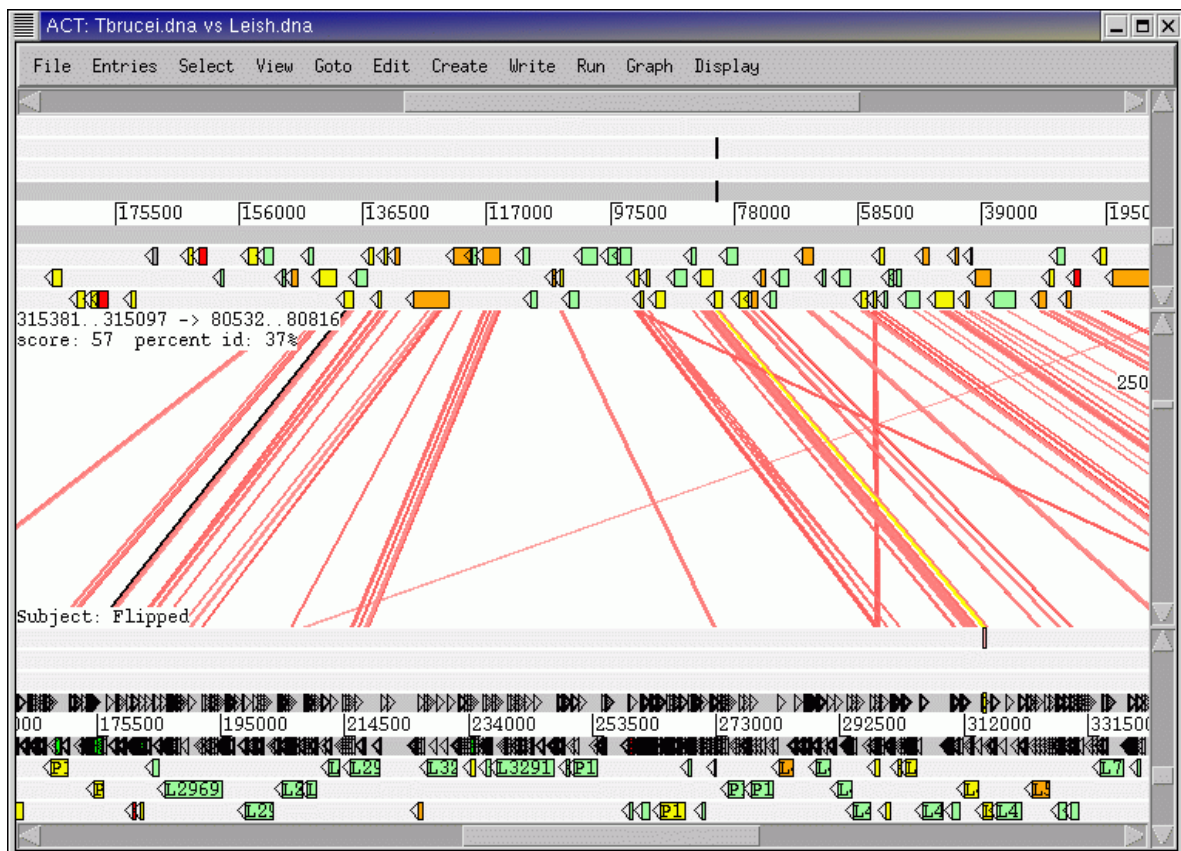
When you have determined where there is synteny, zoom in to the region for a detailed look. At this point you can add the annotation from the files called **Leish.embl** and **Tbrucei.embl**.

Can you see conserved gene order between the 2 species?

Can you see any region where similarity is broken up? Zoom in and look at some of the genes encoded within these regions.

What are the predicted products of the genes assigned to these locations? View the details by clicking on the feature, and then select *'Edit selected feature'* from the *'Edit'* menu after selecting the appropriate CDS feature.

Can you identify any genes in one organism that don't appear to be predicted in the other? If so, add these to your annotation.



## Exercise 4

### Introduction

The quinic acid gene cluster (the *qut* cluster) is present among many filamentous fungi including *Aspergillus fumigatus*, *Neurospora crassa*, *Aspergillus nidulans* and *Podospora anserina*. Although these fungi belong to the same fungal taxonomic family (Ascomycetes), they vary greatly in their biological characteristics. In this exercise you will be studying and comparing the organisation of *qut* gene cluster among these 4 fungi, using ACT.

### Aim

By looking at a comparison of the annotated sequences of *N. crassa*, *A. fumigatus* and *A. nidulans* you will be able to first, add annotations to *qut* cluster genes in *P. anserina* sequence and second compare those genes that are found in all 4 organisms as well as spot the differences and study the synteny.

The files that you are going to need are:

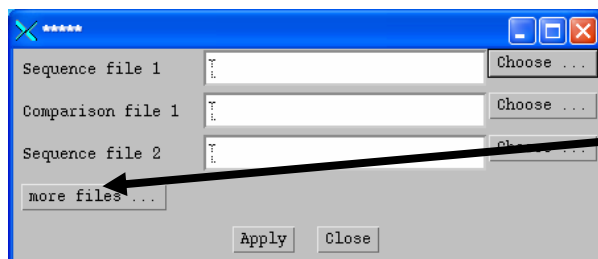
- 1) N\_crassa\_qut.embl - sequence & annotated file for *N. crassa*
- 2) A\_fum\_qut.embl - sequence & annotation file for *A. fumigatus*
- 3) A\_nid\_qut.embl - sequence & annotation file for *A. nidulans* (artificially joined contig)
- 4) P\_anserina\_qut.embl - sequence & gene model file for *P. anserina* (without annotation)
- 5) A\_fum\_N\_crassa.comp - tblastx comparison file of *A. fumigatus* & *N. crassa*
- 6) A\_fum\_A\_nid.comp - tblastx comparison file of *A. fumigatus* & *A. nidulans*
- 7) A\_nid\_P\_anserina.comp - tblastx comparison file of *A. nidulans* & *P. anserina*
- 8) P\_anserina\_N\_crassa.comp - tblastx comparison file of *P. anserina* & *N. crassa*.

First, open an ACT window and then open the annotation and the appropriate comparison files in the order of 1 – 5 – 2 – 6 – 3 – 7 – 4 – 8 – 1 (the numbers are designated above).

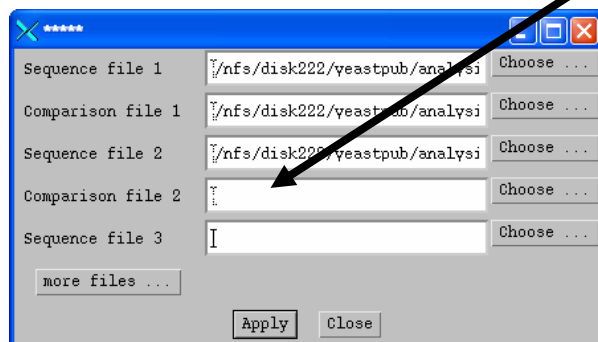
You will need to click on 'more files' to upload more than 2 sequences and the comparison files.

Click on 'apply' after you have uploaded all the files.

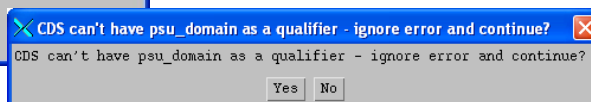
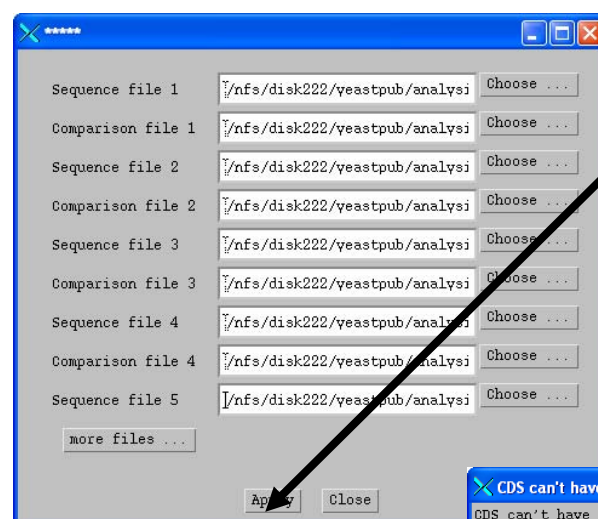
Upload the files in sequential order as described in the previous page



Click on here to load more files and select the appropriate file



Click on here to read all the files that you have selected.



Click on 'yes' if any small dialogue window appears while reading / opening the files.

Can you see any conserved gene order between the *A. fumigatus* & *A. nidulans* in the *qut* gene cluster?

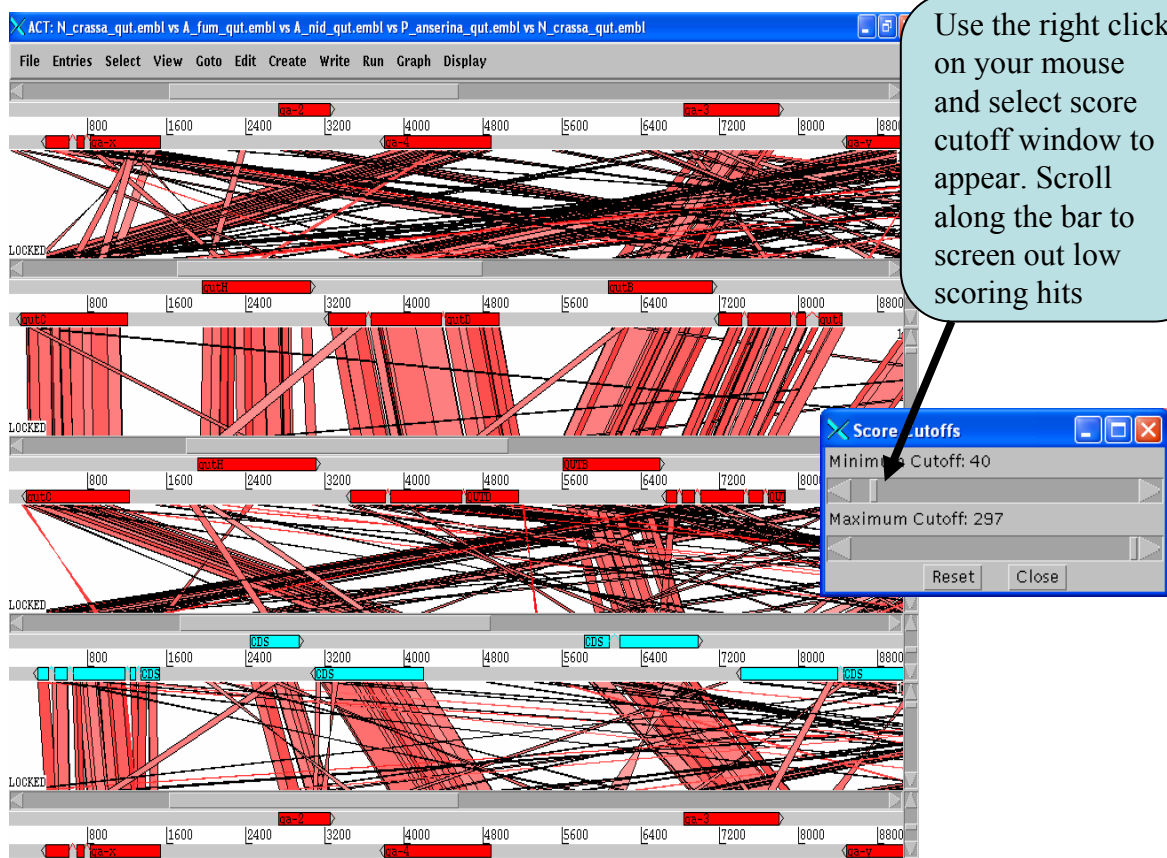
Can you obtain a clearer picture of the ACT 4-way comparison figure by filtering out the low scoring segments, using the blast score cut off feature which you have used previously.

Zoom in and look at some of the genes encoded within these regions. View the details by clicking on the feature, and then select '*Edit selected feature*' from the '*Edit*' menu after selecting the appropriate CDS feature.

By comparing the blast similarity matches, assign your own annotation (gene product) to the predicted gene models (the blue genes) on the *P. anserina* gene model file.

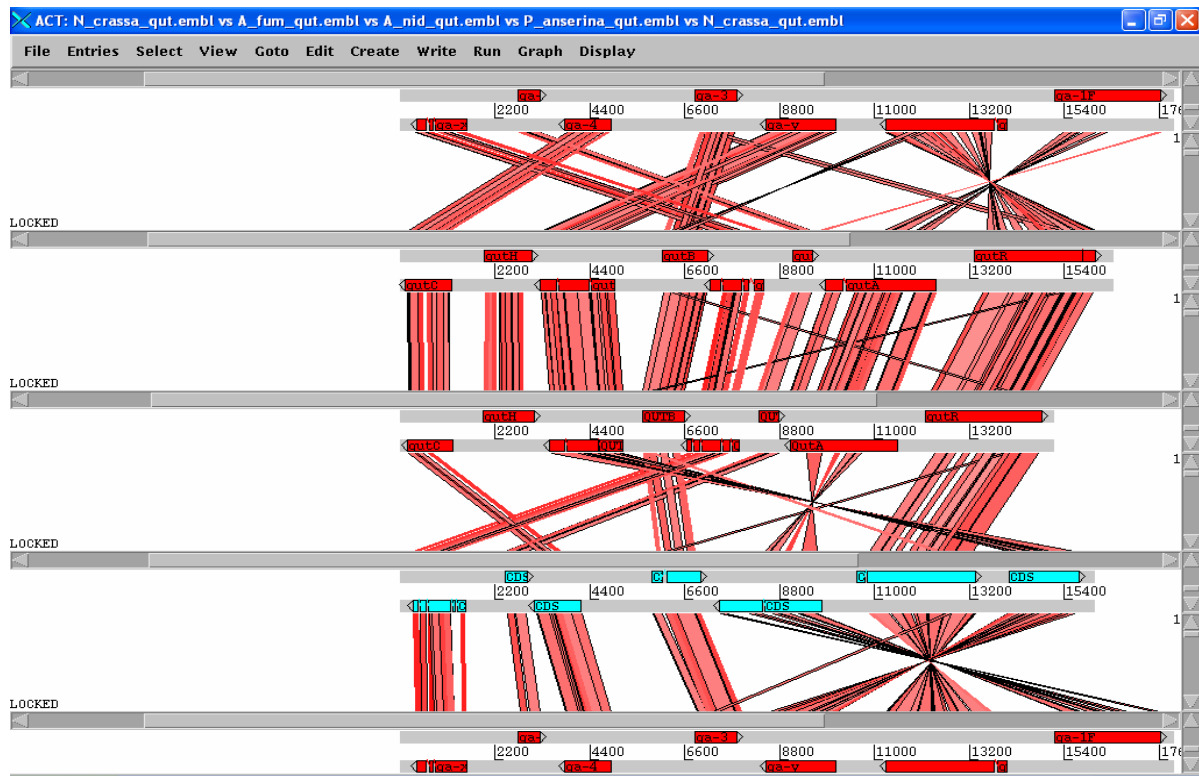
Can you identify any gene NOT present in the *qut* cluster of ALL four fungi?

Note down the gene order (and direction of transcription) in each after you have completed annotation of the *P. anserina* genes in the *qut* cluster.

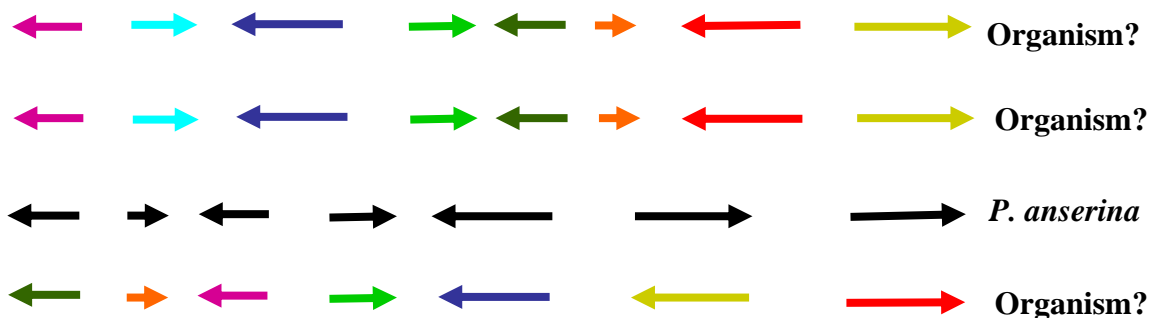




After filtering out the low-scoring blast matches, you should be able to see a figure like the image below.



After comparing the arrangement of genes in the *qut* cluster in these fungi, do you agree with the schematic diagram (not in scale) below where each colour represents a specific type of gene in the quinic acid utilisation gene cluster and each set of clustered genes represents the *qut* cluster one of the organisms. Before you do this you need to annotate the *P. anserina* genes shown as black arrows.



What are these genes?  *qut* ?  *qut* ?  *qut* ?  *qut* ?

*qut* ?  *qut* ?  *qut* ?  *qut* ?

# Module 3

## Generating ACT comparison files using BLAST

### Introduction

In Module 2 you used ACT to visualize pairwise BlastN or TblastX comparisons between DNA sequences. In order to use ACT to investigate your own sequences of interest you will have to generate your own pairwise comparison files. ACT is written so that it will read the output of several different comparison file formats; these are outlined in appendix II. Two of the formats can be generated using Blast software freely downloadable from the NCBI (appendix X). Both Windows and Linux versions of the software are available which can be loaded onto a PC or Mac.

For the purposes of this module the NCBI Blast distribution software has already been installed locally and therefore ready to use. To give you an idea of how easy it is to download and install the software on a PC we have included a step-by-step guide in the appendixes (Appendix X). The example shown in appendix X is for downloading onto a PC with Windows XP. The exercises in this module are based on the Linux version of the Blast software. Although the operating systems are different, the command lines used to run the programs are the same. One of the main differences between the two operating systems is that in Windows the Blast program command line is run in the DOS Command Prompt window, whereas in Linux it is run from a Xterminal window.

### Aims

The aim of this module is to demonstrate how you can generate your own comparison files for ACT from a stand-alone version of the Blast software. In this module you will use Blast to generate comparison files for sequences that you have downloaded from the EBI genomes web resource. A copy of the Blast software has been installed locally. You will run Blast from the command-line using two different programs from the NCBI Blast distribution to generate ACT-readable comparison file for two small sequences (plasmids), and for two large sequences (whole genomes).

### Exercise 1

In this exercise you are going to download two plasmid sequences in EMBL format from the EBI genomes web page. You are then going to use Artemis to write out the DNA sequences of both plasmids in FASTA format. These two FASTA format sequences will then be compared using BlastN to identify regions of DNA-DNA similarity and write out a ACT readable comparison file.

The plasmids chosen for this comparison are the multiple drug resistance incH1 plasmid pHCM1 from the sequenced strain of *Salmonella typhi* CT18 originally isolated in 1993, and R27, another incH1 plasmid first isolated from *S. typhi* in the 1960s.

## Downloading the *S. typhi* plasmid sequences

Go to the EBI genomes web page (<http://www.ebi.ac.uk/genomes/>)

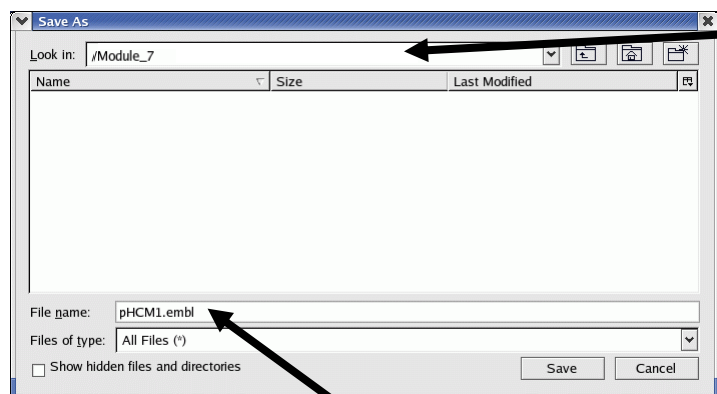
Click on the Plasmid hyperlink

Scroll down the page to the *Salmonella* plasmids

Accession	Description	Length (bp)	Sequence		Proteins
			Plain	HTML	
<b>Acetobacter acetii</b>					
1	Acetobacter acetii pACS	5,123	AF110140	AF110140	2 FASTA SRS
<b>Acetobacter pasteurianus</b>					
2	Acetobacter pasteurianus plasmid pAP12875	1,440	U20550	U20550	2 FASTA SRS
<b>Acidithiobacillus ferrooxidans</b>					
3a	Acidithiobacillus ferrooxidans pTF4.1 plasmid	4,104	X06982	X06982	2 FASTA SRS
3b	Acidithiobacillus ferrooxidans plasmid pTF5	19,792	U73041	U73041	
<b>Acinetobacter sp. EB104</b>					
4	Acinetobacter sp. EB104 plasmid pAC450	4,379	AJ311718	AJ311718	
<b>Acinetobacter pleuropneumoniae</b>					
5	Acinetobacter pleuropneumoniae plasmid pTYM1	4,242	AF203375	AF203375	
<b>Aeromonas salmonicida</b>					
6	Aeromonas salmonicida plasmid pRAS3.2	11,823	AY043299	AY043299	
7	Aeromonas salmonicida subsp. salmonicida plasmid pRAS3.1	11,851	AY043298	AY043298	7 FASTA SRS
<b>Agrobacterium rhizogenes</b>					
8	Agrobacterium rhizogenes plasmid pRi1724	217,594	AF002086	AF002086	169 FASTA SRS
<b>Agrobacterium tumefaciens</b>					
9a	Agrobacterium tumefaciens octopine-type Ti plasmid	194,140	AF242881	AF242881	157 FASTA SRS
9b	Agrobacterium tumefaciens plasmid pTi-SAKURA	206,479	AB016260	AB016260	195 FASTA SRS
10a	Agrobacterium tumefaciens str. C58 (Cereon) plasmid AT (50 parts)	542,869	AE007872	CON	547 FASTA SRS
10b	Agrobacterium tumefaciens str. C58 (Cereon) plasmid TI (20 parts)	214,233	AE007873	CON	197 FASTA SRS
11a	Agrobacterium tumefaciens str. C58 (U. Washington) plasmid AT (49 parts)	542,780	AE008882	CON	543 FASTA SRS

Press the Shift key and left Click on the accession number hyperlink for pHCM1 (AL513383) in the Plain Sequence column

Accession	Organism	Size (bp)	Accession	FASTA SRS
132	Rhodococcus equi plasmid pREAT701	80,610	AF001204	64 FASTA SRS
133	Rhodothermus marinus R-21 plasmid pRM21	2,935	U10426	2 FASTA SRS
134a	Riemerella anatipestifer plasmid pCFC1	3,966	AF048718	4 FASTA SRS
134b	Riemerella anatipestifer plasmid pCFC2	5,609	AF082180	3 FASTA SRS
135	Ruminococcus flavefaciens R13e2 cryptic plasmid pBAW301	1,768	U22411	1 FASTA SRS
136	Salmonella choleraesuis strain 79500 plasmid pSFD10	4,091	AY048853	6 FASTA SRS
137	Salmonella enterica subsp. enterica serovar Berta plasmid pBERT	4,244	AF025795	9 FASTA SRS
138a	Salmonella enterica subsp. enterica serovar Typhi CT18 plasmid pHCM1	218,160	AL513383	234 FASTA SRS
138b	Salmonella enterica subsp. enterica serovar Typhi CT18 plasmid pHCM2	106,516	AL513384	132 FASTA SRS
139a	Salmonella enteritidis serovar Enteritidis plasmid pC	5,249	AY079201	4 FASTA SRS
139b	Salmonella enteritidis serovar Enteritidis plasmid pK	4,245	AY079200	3 FASTA SRS
139c	Salmonella enteritidis serovar Enteritidis plasmid pP	4,301	AY079199	3 FASTA SRS
140a	Salmonella typhi R27 plasmid	180,461	AF250878	207 FASTA SRS
140b	Salmonella typhi plasmid R27	38,245	AF105019	34 FASTA SRS
141	Salmonella typhimurium LT2 strain SGSC1412 plasmid pSLT	93,939	AE006471	102 FASTA SRS
142a	Selenomonas ruminantium pJIM1 plasmid	2,485	Z49917	1 FASTA SRS
142b	Selenomonas ruminantium plasmid pSR1	4,692	AF113972	2 FASTA SRS
143	Shewanella oneidensis MR-1 megaplasmid (15 parts)	161,613	AE014300	125 FASTA SRS
144	Shigella sonnei plasmid ColJs	5,210	AF282884	3 FASTA SRS



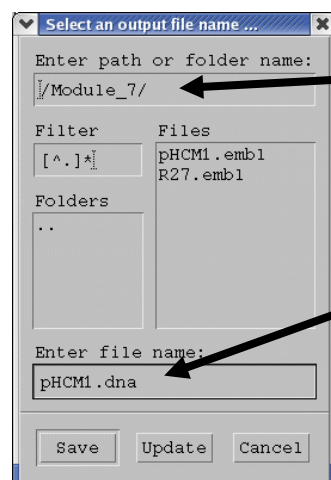
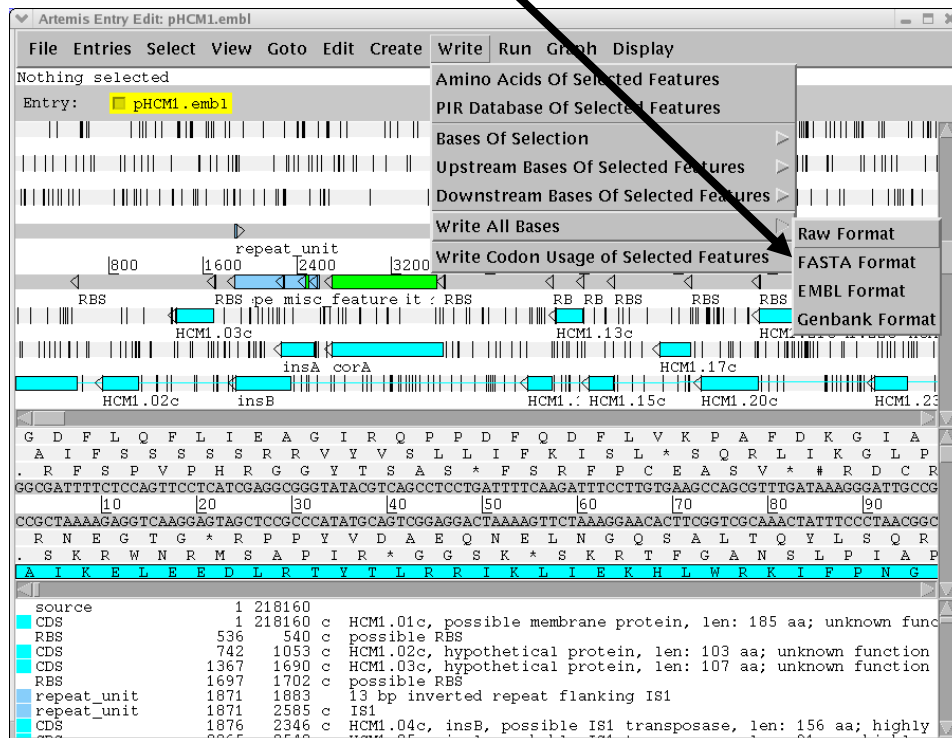
Save the EMBL sequence in the **Module\_7** directory

Save the file as pHCM1.embl

Repeat for the *Salmonella typhi* R27 plasmid (AF250878). Be careful when choosing the plasmid to download as there is also a *Salmonella typhi* plasmid R27 entry (AF105019), the one that you want is the larger of the two, 180,461 kb as opposed to 38,245 kb. Save as R27.embl.

In order to run BlastN you require two DNA sequences in FASTA format. The pHCM1 and R27 sequences previously downloaded from the EBI are EMBL format files, i.e. they contain protein coding information and the DNA sequence. In order to generate the DNA files in FASTA format, Artemis can be used as follows.

Load up the plasmid EMBL files in **Artemis** (each plasmid requires a separate Artemis window), select **Write, Write All Bases, FASTA format**.



Save the DNA sequence in the **Module\_3** directory

Save as pHCM1.dna

Also do this for R27.embl



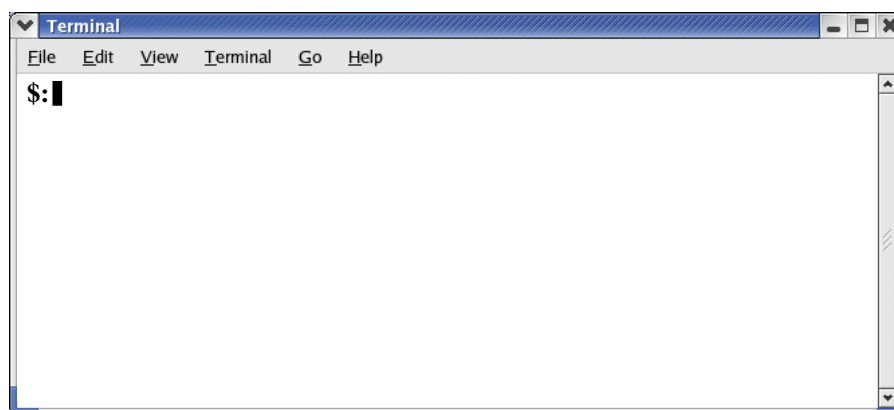
## Running Blast

There are several programs in the Blast package that can be used for generating sequence comparison files. For a detailed description of the uses and options see the appropriate README file in the Blast software directory (see appendix X).

In order to generate comparison files that can be read into ACT you can use the **Blastall** program running either BlastN (DNA-DNA comparison) or TblastX (translated DNA-translated DNA comparison) protocols.

As an example you will run a BlastN comparison on two relatively small sequences; the pHCM1 and R27 plasmids from *S. typhi*. In principle any DNA sequences in FASTA format can be used, although size becomes an issue when dealing with sequences such as whole genomes of several Mb (see exercise 2 in this module). When obtaining nucleotide sequences from databases such as EMBL using a server such as SRS (<http://srs.ebi.ac.uk>), it is possible to specify that the sequences are in FASTA format.

To run the blast software you will need an Xterminal window like the one below. If you do not already have one opened, you can open a new window by clicking on the Xterminal icon on the menu bar at the bottom of your screen.



Make sure you are in the Module\_3 directory. You should now see both the new FASTA files for the pHCM1 and R27 sequences in the Module\_3 directory as well as their respective EMBL format files.

(Hint: You can use the **pwd** command to check the present working directory, the **cd** command to change directories, and the **ls** command will list the contents of the present working directory).

When comparing sequences in Blast, one sequence is designated as a **database** sequence, and the other the **query** sequence. Before you run Blast you have to format one of the sequences so that Blast recognises it as a database sequence. **formatdb** is a program that does this and comes as part of the NCBI Blast distribution.

You will treat pHCM1.dna as the **database** sequence and R27.dna as the **query** sequence

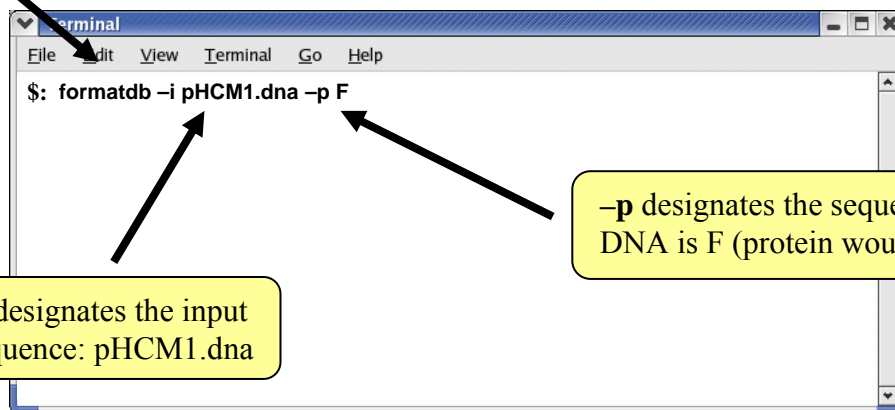
At the Command Prompt type:  
**formatdb -i pHCM1.dna -p F**

Press **Return**

**formatdb** is the  
database format program

**-i** designates the input  
sequence: pHCM1.dna

**-p** designates the sequence type:  
DNA is F (protein would be T)



```
terminal
File Edit View Terminal Go Help
$: formatdb -i pHCM1.dna -p F
```

Now you can run the Blast on the two plasmid sequences. The program that you are going to use is **blastall**. In addition to the standard command line inputs we have to add an additional flag (**-m 8**) to the command line so that the Blast output can be read by ACT. This specifies that the output of Blast is in one line per entry format (see appendix II).

At the Command Prompt type:

**blastall -p blastn -m 8 -d pHCM1.dna -i R27.dna -o pHCM1\_vs\_R27**

Press **Return**

**tblastx** could be substituted here if a  
translated DNA-translated DNA  
comparison was required

**-o** designates the  
output file:  
pHCM1\_vs\_R27

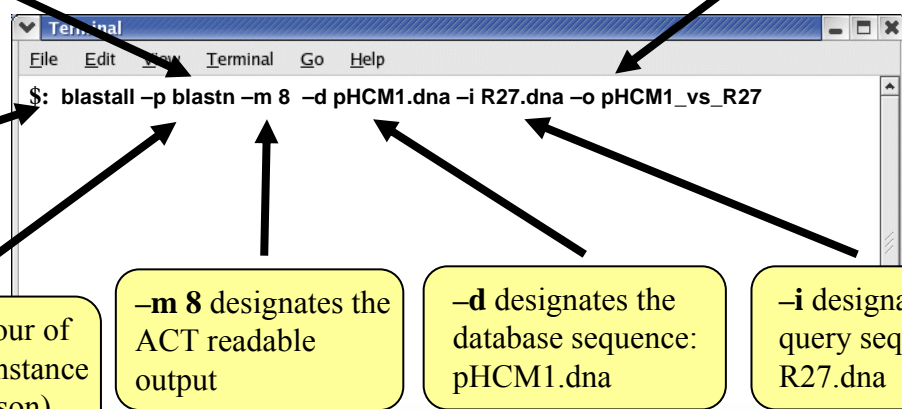
**blastall** is the  
Blast program

**-p** designates the flavour of  
Blast: **blastn** (in this instance  
a DNA-DNA comparison)

**-m 8** designates the  
ACT readable  
output

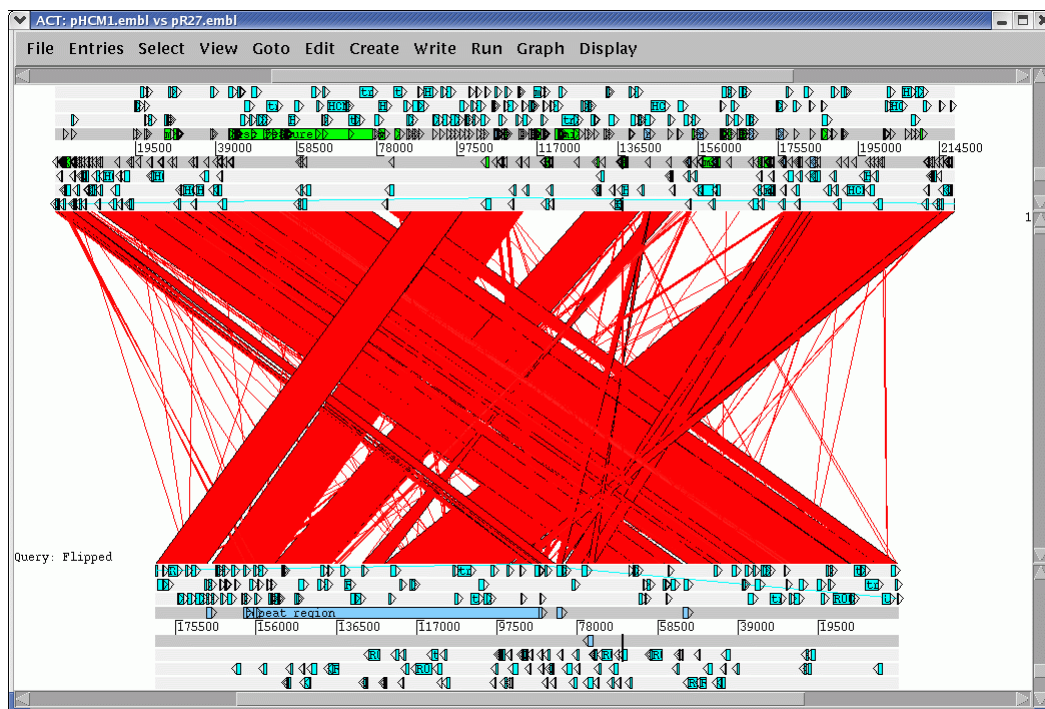
**-d** designates the  
database sequence:  
pHCM1.dna

**-i** designates the  
query sequence:  
R27.dna



```
Terminal
File Edit View Terminal Go Help
$: blastall -p blastn -m 8 -d pHCM1.dna -i R27.dna -o pHCM1_vs_R27
```

The pHCM1\_vs\_R27 comparison file can now be read into ACT along with the pHCM1.embl and R27.embl (or pHCM1.dna and R27.dna) sequence files.



The result of the BlastN comparison shows that there are regions of DNA shared between the plasmids; pHCM1 shares 169 kb of DNA at greater than 99% sequence identity with R27. Much of the additional DNA in the pHCM1 plasmid appears to have been inserted relative to R27 and encodes functions associated with drug resistance. What antibiotic resistance genes can you find in the pHCM1 plasmid that are not found in R27?

The two plasmids were isolated more than 20 years apart. The comparison suggests that there have been several independent acquisition events that are responsible for the multiple drug resistance seen in the more modern *S. typhi* plasmid.

## Exercise 2

In the previous exercise you used BlastN to generate a comparison file for two relatively small sequences (>500,000 kb). In the next exercise we are going to use another program from NCBI Blast distribution, **megablast**, that can be used for nucleotide sequence alignment searches, i.e. DNA-DNA comparisons. If you are comparing large sequences such as whole genomes of several Mb, the **blastall** program is not suitable. The Blast algorithms will struggle with large DNA sequences and therefore the processing time to generate a comparison file will increase dramatically.

**Megablast** uses a different algorithm to Blast which is not as stringent which therefore makes the program faster. This means that it is possible to generate comparison files for genome sequences in a matter of seconds rather than minutes and hours.

There are some drawbacks to using this program. Firstly, only DNA-DNA alignments (BlastN) can be performed using **megablast**, rather than translated DNA-DNA alignments (TblastX) as can be using **blastall**. Secondly as the algorithm used is not as stringent, megablast is suited to comparing sequences with high levels of similarity such as genomes from the same or very closely related species.

In this exercise you are going to download two *Staphylococcus aureus* genome sequences from the EBI genomes web page and use Artemis to write out the FASTA format DNA sequences for both as before in exercise 1. These two FASTA format sequences will then be compared using **megablast** to identify regions of DNA-DNA similarity and write out an ACT readable comparison file.

The genomes that have been chosen for this comparison are from a hospital-acquired methicillin resistant *S. aureus* (MRSA) strain N315 (BA000018), and a community-acquired MRSA strain MW2 (BA000033).

## Downloading the *S. aureus* genomic sequences

Go to the EBI genomes web page (<http://www.ebi.ac.uk/genomes>) as before in exercise 1, and click on the **Bacteria** hyperlink

EMBL-EBI  
European Bioinformatics Institute

Genomes Pages

Genomes Bacteria

Description	Length (bp)	Sequence	Proteins
		Plain HTML	
<b>Agrobacterium tumefaciens</b>			
1a Agrobacterium tumefaciens str. C58 (Cereon) chromosome (circular) (254 parts)	2,841,581	AE0007869 CON	Proteome
1b Agrobacterium tumefaciens str. C58 (Cereon) chromosome (linear) (187 parts)	2,074,782	AE0007870 CON	
2a Agrobacterium tumefaciens str. C58 (U. Washington) chromosome (circular) (256 parts)	2,841,490	AE0006988 CON	Proteome
2b Agrobacterium tumefaciens str. C58 (U. Washington) chromosome (linear) (187 parts)	2,075,560	AE0006989 CON	
<b>Aquifex aeolicus</b>			
3 Aquifex aeolicus VES (109 parts)	1,551,335	AE0006652 CON	Proteome
<b>Bacillus anthracis</b>			
4 Bacillus anthracis str. Ames (18 parts)	5,227,293	AE016879 CON	Proteome
<b>Bacillus cereus</b>			
5 Bacillus cereus ATCC 14579 (18 parts)	5,411,809	AE016877 CON	Proteome
<b>Bacillus halodurans</b>			
6 Bacillus halodurans (14 parts)	4,202,352	BA0000004 CON	Proteome
<b>Bacillus subtilis</b>			
7 Bacillus subtilis subsp. subtilis str. 168 (21 parts)	4,214,630	AI009126 CON	Proteome
<b>Bacteroides thetaiotaomicron</b>			
8 Bacteroides thetaiotaomicron VPI 5482 (21 parts)	6,260,361	AE015928 CON	Proteome
<b>Bifidobacterium longum</b>			
9 Bifidobacterium longum NCC2705 (202 parts)	2,256,646	AE014295 CON	Proteome
<b>Bordetella bronchiseptica</b>			
10 Bordetella bronchiseptica strain RB50 (16 parts)	5,339,179	BX470250 CON	Proteome
<b>Bordetella parapertussis</b>			
11 Bordetella parapertussis strain 13933 (14 parts)	4,779,881	BX470248 CON	Proteome

Scroll down the page to the *Staphylococcus aureus* genomes

Genomes Pages - Bacteria - Mozilla

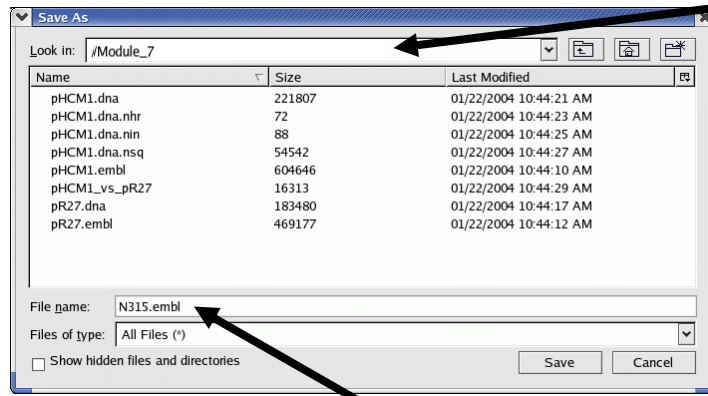
http://www.ebi.ac.uk/genomes/bacteria.html

Genomes Bacteria

87	Rickettsia prowazekii (4 parts)	1,111,523	AI235269 CON	Proteome
<b>Salmonella enterica</b>				
88	Salmonella enterica subsp. enterica serovar Typhi (20 parts)	4,809,037	AI513382 CON	Proteome
89	Salmonella enterica subsp. enterica serovar Typhi Ty2 (16 parts)	4,791,961	AE014613 CON	Proteome
<b>Salmonella typhimurium</b>				
90	Salmonella typhimurium LT2 (220 parts)	4,857,432	AE006468 CON	Proteome
<b>Shewanella oneidensis</b>				
91	Shewanella oneidensis MR-1 (457 parts)	4,969,803	AE014299 CON	Proteome
<b>Shigella flexneri</b>				
92	Shigella flexneri 2a str. 301 (412 parts)	4,607,203	AE005674 CON	Proteome
93	Shigella flexneri 2a str. 2457T (16 parts)	4,599,354	AE014073 CON	3,769 FASTA SRS
<b>Sinorhizobium meliloti</b>				
94	Sinorhizobium meliloti 1021 (12 parts)	3,654,135	AI591688 CON	Proteome
<b>Staphylococcus aureus</b>				
95	Staphylococcus aureus subsp. aureus MW2 (10 parts)	2,820,462	BA0000033 CON	Proteome
96	Staphylococcus aureus subsp. aureus Mu50 (9 parts)	2,878,040	BA0000017 CON	Proteome
97	Staphylococcus aureus subsp. aureus N315 (10 parts)	2,814,000	BA0000018 CON	Proteome
<b>Staphylococcus epidermidis</b>				
	Staphylococcus epidermidis ATCC 12228 (9 parts)	2,499,279	AE015929 CON	Proteome
<b>Staphylococcus agalactiae</b>				
	Staphylococcus agalactiae 2697VR (10 parts)	2,160,267	AE009948 CON	Proteome
	Staphylococcus agalactiae strain 10116 (14 parts)	2,211,485	AI732636 CON	Proteome
<b>Streptococcus mutans</b>				
	Streptococcus mutans UA159 (185 parts)	2,030,921	AE014133 CON	Proteome
<b>Streptococcus pneumoniae</b>				
	Streptococcus pneumoniae R6 (184 parts)	2,038,615	AE007317 CON	Proteome
	Streptococcus pneumoniae TIGR4 (194 parts)	2,160,837	AE003672 CON	Proteome
<b>Streptococcus pyogenes</b>				
	Streptococcus pyogenes M1 GAS (187 parts)	1,852,441	AE004092 CON	Proteome
	Streptococcus pyogenes MGAS315 (37 parts)	1,900,521	AE014074 CON	Proteome
106	Streptococcus pyogenes MGAS232 (173 parts)	1,895,017	AE009949 CON	Proteome
107	Streptococcus pyogenes SSI-1 (6 parts)	1,894,275	BA0000034 CON	Proteome

Press the Shift key and left Click on the *S. aureus* N315 accession number hyperlink (BA000018) in the Plain Sequence column





Save the EMBL sequence in the **Module\_3** directory

Save the file as N315.embl

Repeat for the *S. aureus* MW2 genome (BA000033). Be careful when choosing the genome to download as there is another *S. aureus* genome entry for strain Mu50 (BA000017). Save as MW2.embl.

Generate DNA files in FASTA format using Artemis for both the genome sequences as previously done in exercise 1.

(Hint: In **Artemis** (each genome requires a separate Artemis window), select **Write, Write All Bases, FASTA format**).

Save the DNA sequences as N315.dna and MW2.dna for the respective genomes.

## Running Blast

In the previous exercise you used the **blastall** program to run BlastN on two plasmid sequences. As the genome sequences are larger (~2.8 Mb) you are going to run **megablast**, another program from the NCBI Blast distribution that can generate comparison files in a format that ACT can read (see appendix II). For a detailed description of the uses and options in **megablast** see the megablast README file in the Blast software directory (appendix X).

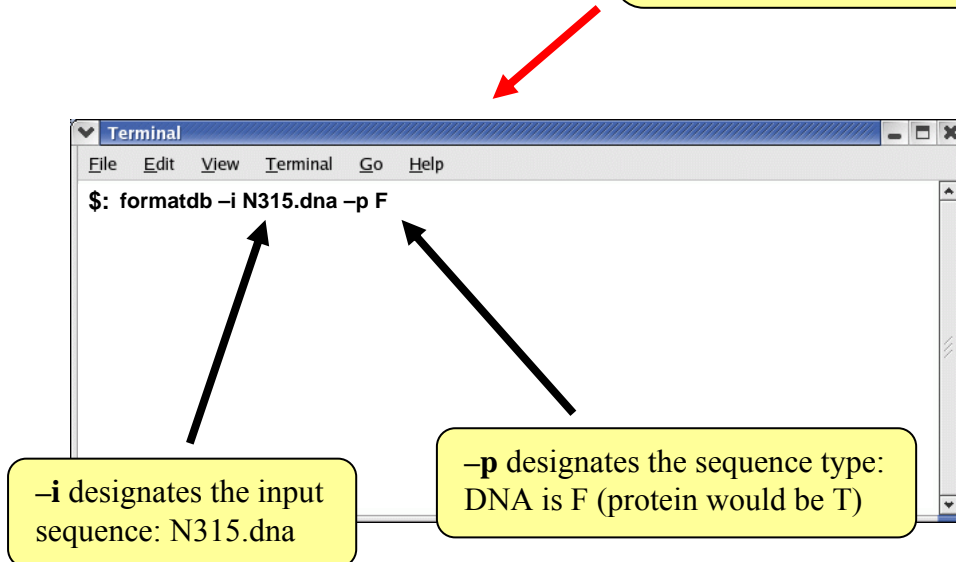
As before you will run the program from the command line in an Xterminal window.

Like Blast, **megablast** requires that one sequence is designated as a **database** sequence and the other the **query** sequence. Therefore one of the sequences has to be formatted so that Blast recognises it as a database sequence. This can be done as before using **formatdb**.

We will treat N315.dna as the **database** sequence and MW2.dna as the **query** sequence

At the Command Prompt type:  
**formatdb -i N315.dna -p F**

Press **Return**



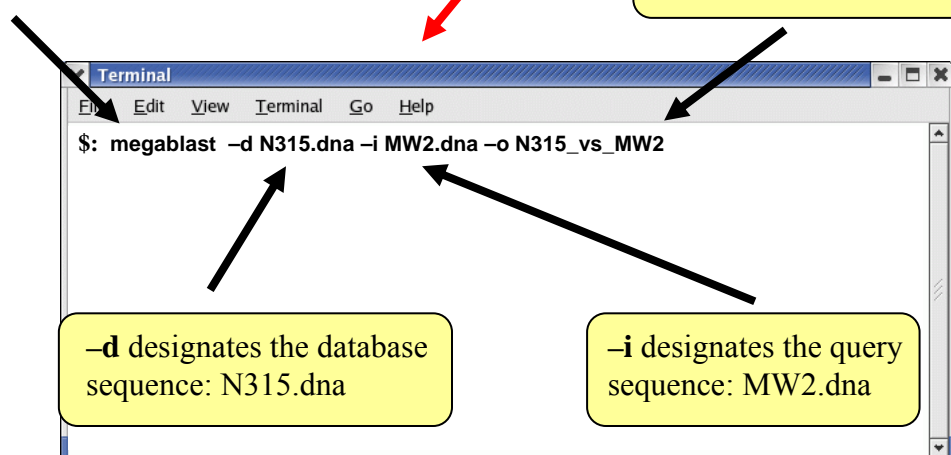
Now we can run the **megablast** on the two MRSA genome sequences. The default output format is one line per entry that ACT can read, therefore there is no need to add an additional flag to the command line (see appendix II).

At the Command Prompt type:  
**megablast -d N315.dna -i MW2.dna -o N315\_vs\_MW2**

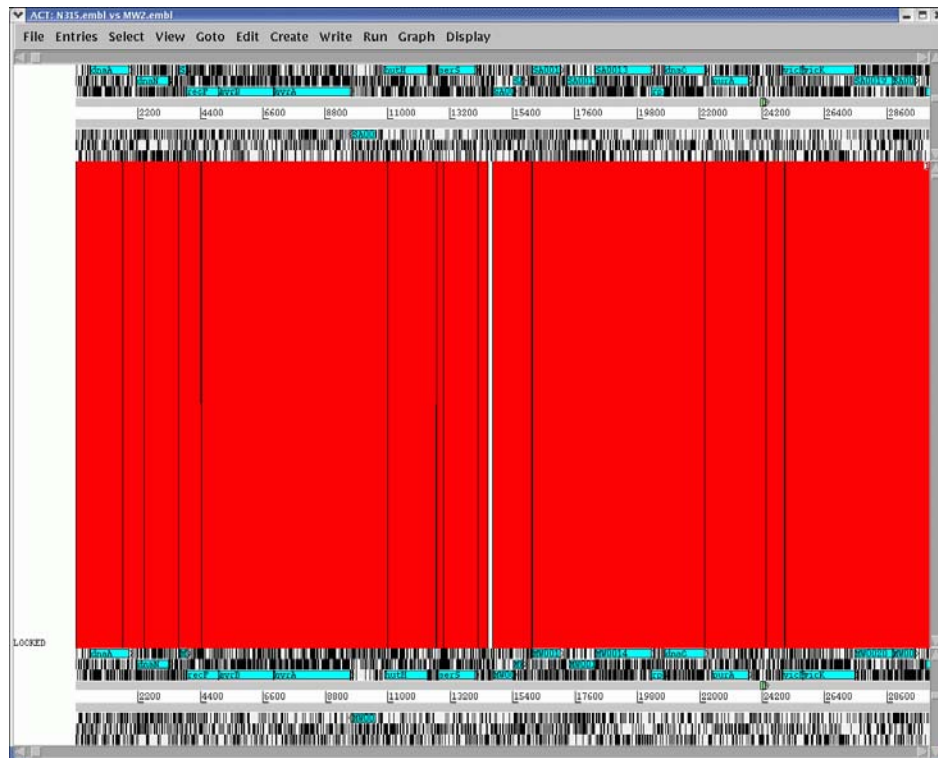
Press **Return**

**megablast** is the program

**-o** designates the output file:  
N315\_vs\_MW2



The N315\_vs\_MW2 comparison file can now be read into ACT along with the N315.embl and MW2.embl (or N315.dna and MW2.dna) sequence files.



A comparison of the N315 and MW2 genomes in ACT using the **megablast** comparison reveals a high level of synteny (conserved gene order). This is perhaps not unsurprising as both genomes belong to strains of the same species. Using results of comparisons like these it is possible to identify genomic differences that may contribute to the biology of the bacteria and also investigate mechanisms of evolution.

Both N315 and MW2 are MRSA, however N315 is associated with disease in hospitals, and MW2 causes disease in the community and is more invasive. Scroll rightward in both genomes to find the first large region of difference. Examine the annotation for the genes in these regions. What are the encoded functions associated with these regions? What significance does this have for the evolution of methicillin resistance in these two *S. aureus* strains from clinically distinct origins?

# Module 4

## Jemboss

# Module 5

# Internet Genome

# Resources

## Introduction

The preceding modules are concerned with predicting genes, and then trying to evaluate what they do. This module will deal firstly with some of the main ways that gene products are described using controlled vocabularies and secondly how you can use these description to quickly access genes from databases.

The module is split into three sections:

### Section 1

#### EC numbers

- a very widely used system for describing enzymes. EC numbers can be used to find out additional information for an enzyme, such as possible orthologues, the biochemical pathway that it's involved in etc, or can be used to identify new enzymes.

### Section 2

#### Gene Ontology

- a way to find genes based on descriptions of the molecular function, biological process or cellular component of their products.

### Section 3

#### InterPro & UniProt

- An integrated documentation resource for protein families, domains and sites

## Aims

The aim of this module will be to explore these controlled vocabularies using a series of worked examples.



## Section 1

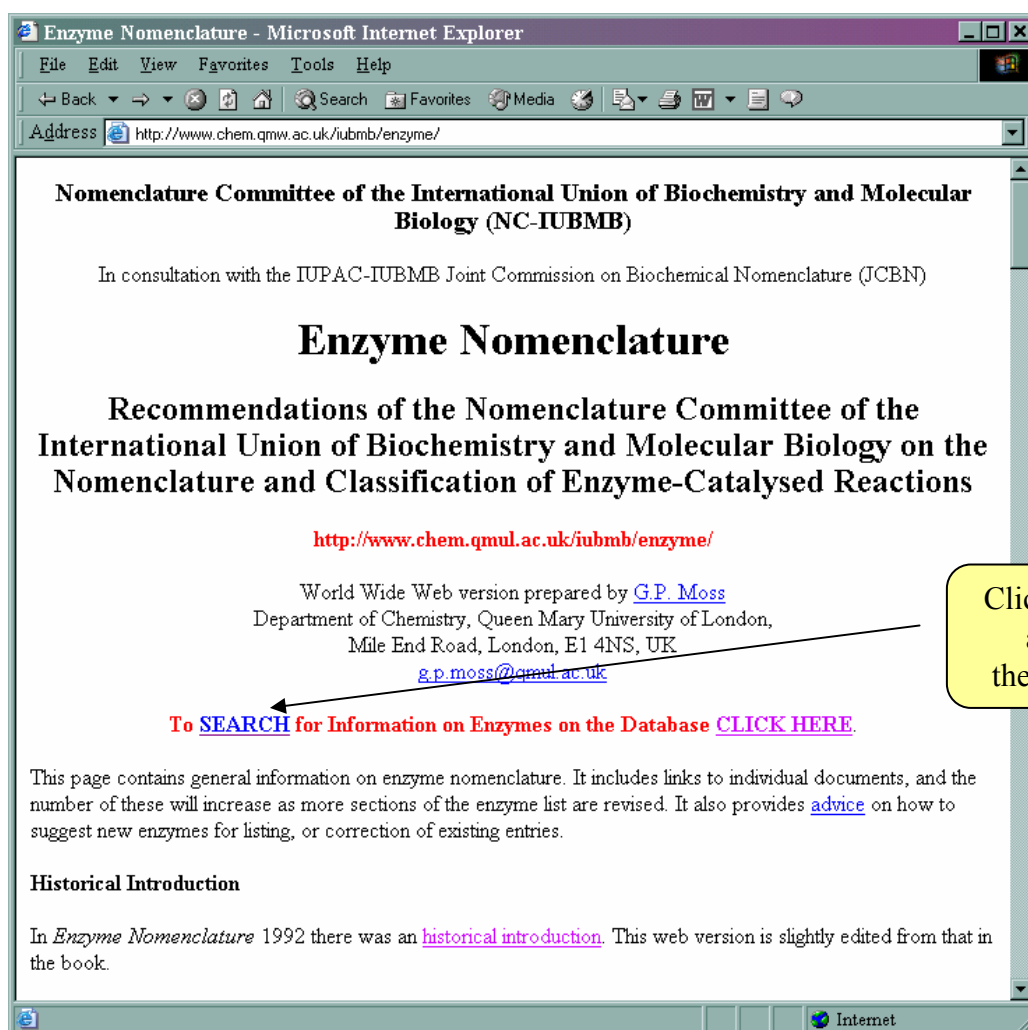
### Exercise 1 Part I

**1. What do *equilase* and 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase do? What kinds of pathways are they involved in?**

You probably won't have a very clear idea of what these enzymes are (even if you're a biochemist).

Use their EC numbers (EC 1.11.1.6 and EC 2.7.7.60, respectively) to find out more from the "official" Enzyme Nomenclature website

Go to this web address <http://www.chem.qmul.ac.uk/iubmb/enzyme/> in your web browser window



## Enter search words

1.11.1.6

All enzymes

Match  words

and return  results

Type in each EC number, and select relevant link from the search results

Each enzyme is represented by a separate web page in IUBMB.

EC 1.11.1.6 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Links

Address <http://www.chem.qmul.ac.uk/iubmb/enzyme/EC1/11/1/6.html>

IUBMB Enzyme Nomenclature

## EC 1.11.1.6

**Common name:** catalase

**Reaction:**  $2 \text{H}_2\text{O}_2 = \text{O}_2 + 2 \text{H}_2\text{O}$

**Other name(s):** equilase; caperase; optidase; catalase-peroxidase; CAT

**Systematic name:** hydrogen-peroxide:hydrogen-peroxide oxidoreductase

**Comments:** A hemoprotein. This enzyme can also act as a peroxidase ([EC 1.11.1.7](#) peroxidase) for which several organic substances, especially ethanol, can act as a hydrogen donor. A manganese protein containing  $\text{Mn}^{\text{III}}$  in the resting state, which also belongs here, is often called pseudocatalase. Enzymes from some microorganisms, such as *Penicillium simplicissimum*, which exhibit both catalase and peroxidase activity, have sometimes been referred to as catalase-peroxidase.

**Links to other databases:** [BRENDA](#), [EXPASY](#), [KEGG](#), [WIT](#), CAS registry number: 9001-05-2

The most commonly used or "official" name is used first

## Exercise 1 Part II

We will now briefly look and explore the other databases listed here. Follow the links shown above.

The BRENDA database contains similar information to the IUBMB site...

The screenshot shows the BRENDA database entry for catalase (EC-Number 1.11.1.6). The browser window title is "BRENDA: Entry of catalase(EC-Number 1.11.1.6) - Mozilla". The address bar shows the URL: [http://www.brenda.uni-koeln.de/php/result\\_flat.php4?ecno=1.11.1.6&organism=](http://www.brenda.uni-koeln.de/php/result_flat.php4?ecno=1.11.1.6&organism=). The BRENDA logo is prominently displayed at the top, with the tagline "The Comprehensive Enzyme Information System". Below the logo, the entry title "Entry of catalase (EC-Number 1.11.1.6)" is shown. A sidebar on the left contains a navigation menu with categories like Enzyme Nomenclature, Enzyme-Ligand Interactions, Functional Parameters, and Molecular Properties. The main content area includes a search bar, a list of organisms to select from (All organism, Acinetobacter sp. (strain ADP1), Ajellomyces capsulata, Anopheles gambiae, Arabidopsis thaliana), and a table of enzyme data. The table has columns for EC NUMBER, COMMENTARY, PATHWAY, and RECOMMENDED NAME. Below the table, there is a section for SYNONYMS and a table of enzyme variants with columns for SYNONYMS, ORGANISM, COMMENTARY, and LITERATURE.

**BRENDA**  
The Comprehensive Enzyme Information System  
Entry of catalase (EC-Number 1.11.1.6)

Any question? -> Use the BRENDA Discussion groups

Mark a special word or phrase in this record:  Mark!

Select one or more organism in this record:  Submit

EC NUMBER	COMMENTARY
1.11.1.6	-

Pathway	KEGG Link
Methane metabolism	<a href="#">00680</a>
Tryptophan metabolism	<a href="#">00380</a>

RECOMMENDED NAME	GeneOntology No.
catalase	<a href="#">4096</a>

**SYSTEMATIC NAME**  
hydrogen-peroxide:hydrogen-peroxide oxidoreductase

SYNONYMS	ORGANISM	COMMENTARY	LITERATURE
caperase	-	-	-
CAT	-	-	-
CatA	Aspergillus nidulans	-	<a href="#">439772</a>
catalase-peroxidase	-	-	-
CatB	Aspergillus nidulans	-	<a href="#">439777</a>
CatF	Pseudomonas syringae	-	<a href="#">439787</a>
equilase	-	-	-
HPI-A	Escherichia coli	catalase-peroxidase isoenzyme	<a href="#">439794</a>
HPI-B	Escherichia coli	catalase-peroxidase isoenzyme	<a href="#">439794</a>
HPH	Escherichia coli	-	<a href="#">439808</a>
HPH	Escherichia coli	monofunctional catalase	<a href="#">439794</a>
KatA	Pseudomonas aeruginosa	-	<a href="#">439780</a>

## EXPASY Database.view

**NiceZyme View of ENZYME: EC [1.11.1.6](#)**

**Official Name**  
Catalase.

**Reaction catalysed**  
 $2 \text{H}_2\text{O}_2 \rightleftharpoons \text{O}_2 + 2 \text{H}_2\text{O}$

**Cofactor(s)**  
Heme; Manganese.

**Comment(s)**

- This enzyme can also act as an EC [1.11.1.7](#) for which several organic substances, especially ethanol, can act as a hydrogen donor.
- A manganese protein containing Mn(3+) in the resting state, which also belongs here, is often called pseudocatalase.
- Enzymes from some microorganisms, such as *Penicillium simplicissimum*, which exhibit both catalase and peroxidase activity, have sometimes been referred to as catalase-peroxidase.

**Human Genetic Disease(s)**  
Acatasia [MIM:115500](#)

**Cross-references**

Biochemical Pathways; map number(s) [K5](#)

PROSITE [PDOC00395](#)

BRENDA [1.11.1.6](#)

PUMA2 [1.11.1.6](#)

PRIAM enzyme-specific profiles [1.11.1.6](#)

Kyoto University LIGAND chemical database [1.11.1.6](#)

IUBMB Enzyme Nomenclature [1.11.1.6](#)

IntEnz [1.11.1.6](#)

MEDLINE [Find literature relating to 1.11.1.6](#)

[P83657](#), CAT1\_COMTR; [Q9C168](#), CAT1\_NEUCR; [P81138](#), CAT1\_PENJA;  
[O8X182](#), CAT2\_NEUCR; [Q9C169](#), CAT3\_NEUCR; [Q96528](#), CATA1\_ARATH;  
[Q27487](#), CATA1\_CAEEL; [P48350](#), CATA1\_CUCPE; [P17598](#), CATA1\_GOSHI;  
[P55307](#), CATA1\_HORVU; [P30264](#), CATA1\_LYCES; [P18122](#), CATA1\_MAIZE;  
[P49315](#), CATA1\_NICPL; [P29611](#), CATA1\_ORYSA; [Q01297](#), CATA1\_RICCO;  
[P49284](#), CATA1\_SOLTU; [P29756](#), CATA1\_SOYBN; [P49319](#), CATA1\_TOBAC;  
[Q43206](#), CATA1\_WHEAT; [P25819](#), CATA2\_ARATH; [Q61235](#), CATA2\_CAEEL;  
[P48351](#), CATA2\_CUCPE; [P30567](#), CATA2\_GOSHI; [P55308](#), CATA2\_HORVU;  
[Q9XHH3](#), CATA2\_LYCES; [P12365](#), CATA2\_MAIZE; [P49316](#), CATA2\_NICPL;  
[P55309](#), CATA2\_ORYSA; [P49318](#), CATA2\_RICCO; [P55312](#), CATA2\_SOLTU;  
[P55313](#), CATA2\_WHEAT; [Q42547](#), CATA3\_ARATH; [P48352](#), CATA3\_CUCPE;  
[P81138](#), CATA3\_NICPL; [Q48560](#), CATA3\_SOYBN.

This link takes you to a digital version of Roche Applied Science “Biochemical Pathways” wall chart

Links to every SwissProt record for this enzyme

## KEGG view of EC 1.11.1.6

ENTRY	EC <a href="#">1.11.1.6</a>
NAME	catalase equilase caperase optidase catalase-peroxidase CAT
CLASS	Oxidoreductases Acting on a peroxide as acceptor Peroxidase
SYSNAME	hydrogen-peroxide:hydrogen-peroxide oxidoreductase
<a href="#">REACTION</a>	<a href="#">2 H2O2 = O2 + 2 H2O</a>
SUBSTRATE	<a href="#">H2O2</a>
PRODUCT	<a href="#">H2O</a>
COFACTOR	<a href="#">O2</a> <a href="#">Heme</a> <a href="#">Manganese</a>
PATHWAY	PATH: <a href="#">MAP00380</a> Tryptophan metabolism PATH: <a href="#">MAP00680</a> Methane metabolism
GENES	HSA: <a href="#">847</a> (CAT) ← MMU: <a href="#">12359</a> (Cas1) RNO: <a href="#">24248</a> (Cat) DME: <a href="#">CG6871</a> (Cat) <a href="#">CG9314</a> CEL: <a href="#">Y54G11A.6</a>

The KEGG database contains tools for analysing the enzymes in pathways

The use of the pathway maps at KEGG will be explored more fully in a later exercise in this module.

## Exercise 1 Part III

Take your web browser back to the IUBMB search page and search using EC2.7.7.60 as before. For some enzymes, you can also get pathway information from their IUBMB pages e.g.

### IUBMB Enzyme Nomenclature

## EC 2.7.7.60

**Common name:** 2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase

**Reaction:** CTP + 2-*C*-methyl-D-erythritol 4-phosphate = diphosphate + 4-(cytidine 5'-diphospho)-2-*C*-methyl-D-erythritol

For diagram [click here](#) ←

**Other name(s):** MEP cytidylyltransferase

**Systematic name:** CTP:2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase

**Comments:** The enzyme from *Escherichia coli* requires Mg<sup>2+</sup> or Mn<sup>2+</sup>. ATP or UTP can replace CTP, but both are less effective. GTP and TTP are not substrates. Forms part of an alternative nonmevalonate pathway for terpenoid biosynthesis (for diagram, [click here](#)).

**Links to other databases:** [BRENDA](#), [EXPASY](#), [KEGG](#), [WIT](#), CAS registry number:

#### References:

- Rohdich, F., Wungsintaweekul, J., Fellermeier, M., Sagner, S., Herz, S., Kis, K., Eisenreich, W., Bacher, A. and Zenk, M.H. Cytidine 5'-triphosphate-dependent biosynthesis of isoprenoids: YgbP protein of *Escherichia coli* catalyzes the formation of 4-diphosphocytidyl-2-*C*-methyl-D-erythritol. *Proc. Natl Acad. Sci. USA* 96 (1999) 11758-11763. [Medline UI: [10518523](#)]
- Kuzuyama, T., Takagi, M., Kaneda, K., Dairi, T. and Seto, H. Formation of 4-(cytidine 5'-diphospho)-2-*C*-methyl-D-erythritol from 2-*C*-methyl-D-erythritol 4-phosphate by 2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase, a new enzyme in the nonmevalonate pathway. *Tetrahedron Lett.* 41 (2000) 703-706.

Click here to get a pathway diagram



## Exercise 2

### 2. What is the EC number of Fructose-1,6-Bisphosphate aldolase

Use the BioCyc Genome Pathway Database to find out the EC number.

<http://www.biocyc.org>

**BioCyc Home** Database Collection

**BioCyc Home Page**

BioCyc is a collection of Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism, with the exception of the MetaCyc database, which is a reference source on metabolic pathways from many organisms. To learn more about BioCyc, read the [Introduction page](#). The BioCyc databases are divided into three tiers, based on their quality.

**BioCyc Databases**

**Tier 1: Intensively Curated Databases**

<a href="#">EcoCyc</a>	<i>Escherichia coli</i> K12
<a href="#">MetaCyc</a>	Metabolic pathways and enzymes from 300 organisms
<a href="#">BioCyc Open Chemical Database</a>	Chemical compound database

**Tier 2: Computationally-Derived Databases Subject to Moderate Curation**

17 databases are available. [\[list of tier 2 DBs\]](#)

**Tier 3: Computationally-Derived Databases Subject to No Curation**

142 databases are available and ready for adoption [\[more\]](#) by interested scientists for curation and updating. PGDBs in Tier 3 were produced as a collaboration between the groups of Peter D. Karp at SRI International and Christos Ouzounis at the European Bioinformatics Institute. [\[list of tier 3 DBs\]](#)

**Other Pathway/Genome Databases on the Internet**

<a href="#">AraCyc</a>	<i>Arabidopsis thaliana</i>	S. Rhee, <a href="#">Department of Plant Biology, Carnegie Institution, USA</a>
<a href="#">LacPlantCyc</a>	<i>Lactobacillus plantarum</i> WCFS1	F. H. J. van Enckevort, <a href="#">CMBI</a> , The Netherlands
<a href="#">MjCyc</a>	<i>Methanococcus jannaschii</i>	C. Ouzounis, <a href="#">European Bioinformatics Institute</a> , UK
<a href="#">PseudoCyc</a>	<i>Pseudomonas aeruginosa</i>	F. Brinkman, <a href="#">Pseudomonas Genome Project</a> , Simon Fraser U., Canada
<a href="#">Yeast Biochemical Pathways</a>	<i>Saccharomyces cerevisiae</i>	<a href="#">SGD</a> curators, Stanford U., USA

[Contact us](#) if you'd like your PGDB added to this list.

**Registry of Downloadable Pathway/Genome Databases**

**BioCyc Home**

**Search**  
[Database Search](#)  
[Advanced Database Search](#)  
[Help](#)

**News**  
 Jul 11 [BioCyc grows to 160 DBs](#)  
 May 23 [BioCyc 9.1](#)

**Services**  
[Software/Data Download](#) including:  
 BioPAX format  
 SBML format  
[User Support](#)  
[Subscribe to Mailing List](#)  
[EcoCyc T-shirts](#)

**Information**  
[Introduction to BioCyc](#)  
[162 Pathway/Genome DBs](#)  
[Guided Tour](#)  
[Pathway Tools Software](#)  
[Publications](#)  
[Linking to BioCyc](#)  
[External Links](#)

Search text of BioCyc web pages:  
  
 Powered by Google

**Click here to enter**

**BioCyc Query Page**

[Tutorial: Creating a Pathway/Genome Database March 15 - 18, 2004](#)  
[Click here for more info.](#)

This form provides several different mechanisms for querying Pathway/Genome Databases.

**Select a dataset:**

• **Query**

To retrieve objects by name, first select the type of object you wish to retrieve, then enter the name and click Submit. All objects containing that name as a substring will be returned. You may also enter names or EC numbers, separating them with commas.

• **Choose from a [list of pathways](#)**

• **Browse Ontology:**

Each dataset contains classification hierarchies for pathways, for reactions (the enzyme nomenclature system), for compounds, and for genes. Select a classification system to browse.

• **Links to summary information about the selected organism:**

- [Summary page for dataset](#)
- [Metabolic Overview Diagram](#) [Expression Viewer](#) (not available for MetaCyc)

Enter product name here and hit submit

**Query Results**

The query **aldolase** matched the following objects:

**Proteins**

- [2-dehydro-3-deoxyphosphogalactonate aldolase](#)
- [2-dehydro-3-deoxyphosphoheptonate aldolase \(protein complex\)](#)
- [2-dehydro-3-deoxyphosphoheptonate aldolase \(protein complex\) \(protein complex\)](#)
- [2-dehydro-3-deoxyphosphoheptonate aldolase \(protein complex\) \(protein complex\)](#)
- [2-keto-3-deoxy-6-phosphogluconate aldolase](#)
- [3-deoxy-D-manno-octulosonic acid 8-phosphate synthase \(2-dehydro-3-deoxyphosphooctonate aldolase\)](#)
- [4-hydroxy-2-ketovalerate aldolase](#)
- [alpha-dehydro-beta-deoxy-D-glucarate aldolase](#)
- [citrate lyase \(citrate aldolase\)](#)
- [deoxyribose-phosphate aldolase](#)
- [dihydroneopterin aldolase](#)
- [fructose 6-phosphate aldolase 1](#)
- [fructose 6-phosphate aldolase 2](#)
- [fructose biphosphate aldolase class I](#)
- [fructose biphosphate aldolase class II](#)
- [fructose biphosphate aldolase monomer \(polypeptide\)](#)
- [fructose biphosphate aldolase monomer \(polypeptide\)](#)
- [glycine hydroxymethyltransferase \(serine aldolase\)](#)
- [L-allo-threonine aldolase](#)
- [L-fucose-phosphate aldolase](#)
- [N-acetylneuraminic acid aldolase](#)
- [protein with similarity to HHED aldolases](#)

Select this enzyme

**fructose biphosphate aldolase class I - Microsoft Internet Explorer**

Address: <http://biocyc.org:1555/ECOLI/NEW-IMAGE?type=ENZYME&object=FRUCBISALD-CLASSI>

## *E. coli* K-12 Enzyme: fructose biphosphate aldolase class I

Superclasses: [Complexes](#) -> [Protein-Complexes](#)  
[Proteins](#) -> [Protein-Complexes](#)

Comment:  
 The typical class I aldolases of plants and animals have been thoroughly studied [Baldwin78]. Fructose-1,6-bisphosphate aldolases can be divided into two classes on the basis of their catalytic and structural properties. [Baldwin78] Class I fructose 1,6 biphosphate aldolases were once thought to be confined to eukaryotic organisms but have since been detected in several bacterial species. [Baldwin78a] The occurrence of such an aldolase in bacteria was unexpected in light of the distribution of aldolases. [Stribling73] The enzymes of eukaryotes generally fall into Class I and are tetramers of polypeptide chains. [Alefounder89a] In earlier studies [Stribling73] it was thought that the class I *E. coli* aldolase was tetrameric with a mol. wt. of approx. 140K. In 1978 new purification techniques were used. The true aldolase could be measured by using Fru-1,6-P<sub>2</sub> that had been purified by chromatography on DEAE-cellulose to remove the fructose-6-phosphate. Using these methods the enzyme appeared to be larger than was previously supposed and may be a decamer with a mol. wt. of approx. 340,000. The size of aldolase 1 and the effect of cross-linking reagents on it, indicate that its structure must differ significantly from that of the typical tetrameric class-I enzymes from eukaryotes. [Baldwin78a, Stribling73.]

Component Composition: [fructose biphosphate aldolase monomer x 10](#)

Species: [Escherichia coli K-12](#)

Gene-Reaction Schematic: [4.1.2.13](#)

Make a note of the EC number

Clicking on the thumbnail map links to more detail

### Exercise 3

3. Here are six EC numbers for proteins in the malaria genome annotation that have been assigned based on protein similarities. Are they involved in a common pathway? If so, can you use KEGG to piece together the pathway and predict which gene is missing and therefore could remain unidentified in the Malaria genome.

The first EC number is for a fructose biphosphate aldolase and you already have it (see previous exercise).

The remaining known EC numbers are listed below:

2.7.1.1  
 5.4.2.8  
 2.7.7.13  
 4.2.1.47  
 2.7.1.90

Next, use KEGG to find possible pathways in which all these enzymes are found.  
<http://www.genome.jp/kegg>

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG Anniversary Symposium  
December 15-16, 2005, Kyoto, Japan

## KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and chemical information. Towards this end we have been developing a bioinformatics resource named KEGG, Kyoto Encyclopedia of Genes and Genomes, as part of the research projects in the Kanehisa Laboratory of Kyoto University Bioinformatics Center.

- Main entry point to the KEGG web service**  
[KEGG2](#)    [KEGG Table of Contents](#)
- Four constituent databases of KEGG**
  - PATHWAY** 30,417 pathways generated from 242 reference pathways
  - GENES** 1,099,188 genes in 31 eukaryotes + 233 bacteria + 23 archaea
  - LIGAND** 12,849 compounds, 2,435 drugs, 11,152 glycans, 6,448 reactions
  - BRITE** 7,351 KO (KEGG Orthology) groups
- Selected KEGG Organisms**
  - [hsa](#) (human) [mmu](#) (mouse) [rno](#) (rat) [dre](#) (zebra fish) [dme](#) (fruit fly) [cel](#) (nematode)
  - [ath](#) (thale cress) [sce](#) (budding yeast) [spo](#) (fission yeast) [eco](#) (E. coli) [bsu](#) (B. subtilis)
- Quick search by DBGET**  
 Search  for     
 (Example) [Alzheimer](#)

<b>Introduction</b>	<b>Standards</b>	<b>Links</b>	<b>GenomeNet</b>
<a href="#">User manuals</a>	<a href="#">XML</a>	<a href="#">Related databases</a>	<a href="#">FTP access</a>
<a href="#">References</a>	<a href="#">API</a>		<a href="#">KegDraw/KegArray</a>
<b>Feedback</b>	<b>Release notes</b>	<b>Disclaimer</b>	
	KEGG Release 35.1, September 2005 (plus daily updates)		

Copyright 1995-2005 Kanehisa Laboratories




KEGG Encyclopedia - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.genome.jp/kegg/kegg2.html

Home Bookmarks Google Library SSC SSC-dev WIKI PSORT Pfam SRS7 FTP E.coli DB GeneDB Entrez PubMed

 **KEGG - Table of Contents**

KEGG2 PATHWAY GENES LIGAND BRITE XML API DBGET

### Generalized KEGG

Content	Database	Search & Compute	DBGET Search
Pathway information	KEGG PATHWAY	Search objects in KEGG pathways Color objects in KEGG pathways KEGG pathways in XML	PATHWAY
Genomic information	KEGG GENES	BLAST search against GENES/GENOME FASTA search against GENES/GENOME KEGG EXPRESSION	GENES DGENES / EGENES GENOME
Chemical information	KEGG LIGAND	Search similar compound structures Search similar glycan structures Predict reactions and assign EC numbers Generate possible reaction paths	COMPOUND DRUG GLYCAN REACTION RPAIR ENZYME LIGAND
Binary relations and hierarchies	KEGG BRITE	KEGG Orthology (KO) Automatic annotation (KO assignment) Therapeutic category of drugs	KO

**Specialized KEGG**

**KEGG for specific organisms**

Enter KEGG organism code:    (examples) hsa mmu sce eco bsu syn

Or use the list of [KEGG organisms](#)

Customize the organism menu with selected organisms

Show currently selected organisms (All organisms in GENES)

**KEGG for selected research areas**

[KEGG ANNOTATION](#) for genome/EST annotation by KAAS  
[KEGG EXPRESSION](#) for microarray data analysis  
[KEGG GLYCAN](#) for glycome informatics

**Click here**

Other options include allowing specific enzymes in a pathway to be colour coded

Search Objects in KEGG Pathways - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.genome.jp/kegg/tool/search\_pathway.html

Search

Home Bookmarks Google Library SSC SSC-dev WIKI PSORT Pfam SRS7 FTP E.coli DB GeneDB Entrez PubMed

**Search Objects in KEGG Pathways**

Search against: Reference pathway

Enter objects:

(Examples for Reference pathway)  
ec:5.3.1.1 cpd:C00111 cpd:C00118  
1.2.1.12 C00236

(Examples for Homo sapiens pathway)  
hsa:7167 hsa:GPI ec:2.7.1.11 cpd:C00118  
ALDOA 1.2.1.12 C00236

Alternatively, enter the file name containing the data:

Browse...

☐ Display objects NOT found in the search

Exec Clear

[ KEGG2 | KEGG | GenomeNet ]

1 Type in or paste your list of EC numbers here

2 Press exe

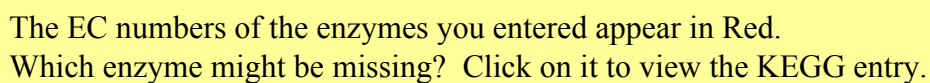
In which pathway do all the enzymes from the list act?

Click on the map for that pathway. See over page.

### Pathway Search Result

Click here

- ◆ [map00010 Glycolysis / Gluconeogenesis](#)  
EC 2.7.1.1 hexokinase; hexokinase type IV glucokinase; hexokinase D; hexokinase type IV; hexokinase (phosphorylating); ATP-dependent h
- ◆ [map00051 Fructose and mannose metabolism](#)  
EC 2.7.1.1 hexokinase; hexokinase type IV glucokinase; hexokinase D; hexokinase type IV; hexokinase (phosphorylating); ATP-dependent h  
EC 2.7.1.90 diphosphate-fructose-6-phosphate 1-phosphotransferase; 6-phosphofructokinase (pyrophosphate); pyrophosphate-fructose 6-phos  
EC 2.7.7.13 mannose-1-phosphate guanylyltransferase; GTP-mannose-1-phosphate guanylyltransferase; PIM-GMP (phosphomannose isomerase-gua  
EC 4.2.1.47 GDP-mannose 4,6-dehydratase; guanosine 5'-diphosphate-D-mannose oxidoreductase; guanosine diphosphomannose oxidoreductase;  
EC 5.4.2.8 phosphomannomutase; mannose phosphomutase; phosphomannose mutase; D-mannose 1,6-phosphomutase
- ◆ [map00052 Galactose metabolism](#)  
EC 2.7.1.1 hexokinase; hexokinase type IV glucokinase; hexokinase D; hexokinase type IV; hexokinase (phosphorylating); ATP-dependent h
- ◆ [map00500 Starch and sucrose metabolism](#)  
EC 2.7.1.1 hexokinase; hexokinase type IV glucokinase; hexokinase D; hexokinase type IV; hexokinase (phosphorylating); ATP-dependent h
- ◆ [map00521 Streptomycin biosynthesis](#)  
EC 2.7.1.1 hexokinase; hexokinase type IV glucokinase; hexokinase D; hexokinase type IV; hexokinase (phosphorylating); ATP-dependent h
- ◆ [map00530 Aminosugars metabolism](#)  
EC 2.7.1.1 hexokinase; hexokinase type IV glucokinase; hexokinase D; hexokinase type IV; hexokinase (phosphorylating); ATP-dependent h



-74-

## Section 2

### Gene Ontology (GO)

The official browser for GO annotations is "Amigo".

Web address: <http://www.godatabase.org/>

The structure of the ontology allows you to quickly find the description or term that you're looking for. The 'tree' of GO terms describes various facts of the proteins, function, cellular location etc. AmiGO holds all the gene predictions from multiple genomes within this tree. First, navigate through the "tree" by expanding and collapsing branches...

**AmiGO**

Search GO

☐ Exact Match

☒ Terms

☐ Gene Symbol/Name

Advanced Query

Query By Sequence

**Gene Product Filters**

**Species**

All

A. aeolicus

A. fulgidus

**Datasource**

All

FlyBase

SGD

**Evidence Code**

All Curator Approved

IC

IMP

XML

Flat File

Permalink

Last updated: 2005-09-06

**all : all (219618)**

- GO:0008150 : biological\_process ( 145845 )
- GO:0005575 : cellular\_component ( 131834 )
- GO:0003674 : molecular\_function ( 152664 )
- obsolete\_biological\_process : obsolete\_biological\_process ( 68 )
- obsolete\_cellular\_component : obsolete\_cellular\_component ( 27 )
- obsolete\_molecular\_function : obsolete\_molecular\_function ( 773 )

**Graphical View**

Change the datasource in Amigo tool bar to filtering by organism

Click to expand branches

Here's an expanded view of biological process and cellular component.

**all : all (219618)**

- GO:0008150 : biological\_process ( 145845 )
  - GO:0007610 : behavior ( 4698 )
    - GO:0000004 : biological process unknown ( 36389 )
  - GO:0009987 : cellular process ( 91982 )
  - GO:0007275 : development ( 18235 )
  - GO:0040007 : growth ( 3622 )
  - GO:0007582 : physiological process ( 97174 )
  - GO:0043473 : pigmentation ( 174 )
  - GO:0050789 : regulation of biological process ( 19190 )
  - GO:0000003 : reproduction ( 5336 )
  - GO:0016032 : viral life cycle ( 306 )
- GO:0005575 : cellular\_component ( 131834 )
  - GO:0005623 : cell ( 96315 )
    - GO:0008372 : cellular component unknown ( 28690 )
  - GO:0031012 : extracellular matrix ( 1062 )
  - GO:0005576 : extracellular region ( 9945 )
  - GO:0043226 : organelle ( 66478 )
  - GO:0043234 : protein complex ( 13971 )
  - GO:0019012 : virion ( 133 )
- GO:0003674 : molecular\_function ( 152664 )
- obsolete\_biological\_process : obsolete\_biological\_process ( 68 )
- obsolete\_cellular\_component : obsolete\_cellular\_component ( 27 )
- obsolete\_molecular\_function : obsolete\_molecular\_function ( 773 )

Two different symbols were used to denote how terms are related to each other

The numbers indicate the number of genes with this function



## Exercise 1

Next use the "search GO" box search for the term "DNA helicase" and submit. several GO terms will appear.

The screenshot shows the AmiGO web interface in a Mozilla browser window. The search bar at the top contains 'DNA helicase'. The search results are displayed in a table with columns: GO Term, GO ID, Match Synonym, Ontology, Definition, and Comment. The first result is 'DNA helicase activity' (GO:0003678). A yellow callout box with the text 'Click here to show the term in the GO tree' points to the 'DNA helicase activity' term in the first row.

GO Term	GO ID	Match Synonym	Ontology	Definition	Comment
<a href="#">DNA helicase activity</a>	GO:0003678		F	Catalysis of the hydrolysis of ATP to unwind the DNA helix at the replication fork, allowing the resulting single strands to be copied.	Consider also annotating to the molecular function term 'DNA binding'; <a href="#">GO:0003677</a> .
<a href="#">ATP-dependent DNA helicase activity</a>	GO:0004003		F	Catalysis of the reaction: ATP + H <sub>2</sub> O = ADP + phosphate, driving the unwinding of the DNA helix.	Consider also annotating to the molecular function term 'ATP binding'; <a href="#">GO:0005524</a> .
<a href="#">DNA helicase IV activity</a>	GO:0008722		F		
<a href="#">single-stranded DNA-dependent ATP-dependent DNA helicase activity</a>	GO:0017116		F	Catalysis of the reaction: ATP + H <sub>2</sub> O = ADP + phosphate, in the presence of single-stranded DNA, driving the unwinding of a DNA helix.	Consider also annotating to the molecular function term 'ATP binding'; <a href="#">GO:0005524</a> .
<a href="#">single-stranded DNA-dependent ATP-dependent DNA helicase complex</a>	GO:0017117		C		
<a href="#">3' to 5' DNA helicase activity</a>	GO:0043138		F	Catalysis of the unwinding of the DNA helix in the direction 3' to 5'.	
<a href="#">5' to 3' DNA helicase activity</a>	GO:0043139		F	Catalysis of the unwinding of the DNA helix in the direction 5' to 3'.	
<a href="#">ATP-dependent 3' to 5' DNA helicase activity</a>	GO:0043140		F	Catalysis of the reaction: ATP + H <sub>2</sub> O = ADP + phosphate, driving the unwinding of the DNA helix in the direction 3' to 5'.	Consider also annotating to the molecular function term 'ATP binding'; <a href="#">GO:0005524</a> .
<a href="#">ATP-dependent 5' to 3' DNA helicase activity</a>	GO:0043141		F	Catalysis of the reaction: ATP + H <sub>2</sub> O = ADP + phosphate, driving the unwinding of the DNA helix in the direction 5' to 3'.	Consider also annotating to the molecular function term 'ATP binding'; <a href="#">GO:0005524</a> .

At the bottom of the table, there is a dropdown menu set to 'Show checked items in tree', a 'Check/Uncheck All' button, and a 'Submit Query' button.

Footer text: Help GOst The Gene Ontology GO Request AmiGO Request  
Last updated: 2005-09-06  
Copyright The Gene Ontology Consortium



## Exercise II

By selecting an individual datasource, see how many *Vibrio cholerae* sequences have been annotated (see over page)

The screenshot shows the AmiGO web interface. On the left, there is a search bar with 'DNA helicase' entered. Below it are filters for Species (T. maritima, T. volcanium, V. cholerae), Datasource (Ensembl, RGD, TIGR\_CMR), Evidence Code (All Curator Approved, IC, IMP), and a 'Set Filters' button. A 'Query Summary' box shows 'Your Query: DNA helicase', 'Exact Match: no', 'Target: Terms', 'Fields: Name and Synonyms', and 'Results: 9'. The main table lists results with columns: GO Term, GO ID, Match Synonym, Ontology, Definition, and Comment. An arrow points from the 'Datasources' label to the 'TIGR\_CMR' filter.

GO Term	GO ID	Match Synonym	Ontology	Definition	Comment
DNA helicase activity	GO:0003678		F	Catalysis of the hydrolysis of ATP to unwind the DNA helix at the replication fork, allowing the resulting single strands to be copied.	Consider also annotating to the molecular function term 'DNA binding'; GO:0003677.
ATP-dependent DNA helicase activity	GO:0004003		F	Catalysis of the reaction: ATP + H <sub>2</sub> O = ADP + phosphate, driving the unwinding of the DNA helix.	Consider also annotating to the molecular function term 'ATP binding'; GO:0005524.
DNA helicase IV activity	GO:0008722		F		
single-stranded DNA-dependent ATP-dependent DNA helicase activity	GO:0017116		F	Catalysis of the reaction: ATP + H <sub>2</sub> O = ADP + phosphate, in the presence of single-stranded DNA, driving the unwinding of a DNA helix.	Consider also annotating to the molecular function term 'ATP binding'; GO:0005524.
single-stranded DNA-dependent ATP-dependent DNA helicase complex	GO:0017117		C		
3' to 5' DNA helicase activity	GO:0043138		F	Catalysis of the unwinding of the DNA helix in the direction 3' to 5'.	
5' to 3' DNA helicase activity	GO:0043139		F	Catalysis of the unwinding of the DNA helix in the direction 5' to 3'.	

The screenshot shows the AmiGO web interface with a search for 'DNA helicase'. The 'Datasource' filter is set to 'TIGR\_CMR'. The results are listed as a tree structure. A callout box points to the 'GO:0003678 : DNA helicase activity (166)' term, stating: 'Click on the term to find individual DNA helicases.' Another callout box points to the 'Datasource' filter, stating: 'Vibrio cholerae annotation is from the CMR database at TIGR'.

**Search GO:**  ☐ Exact Match ☒ Terms ☐ Gene Products

**Datasource:**

**GO:0003678 : DNA helicase activity (166)**

- GO:0008150 : biological process (72641)
- GO:0005575 : cellular component (59242)
- GO:0003674 : molecular function (94552)
- GO:0003824 : catalytic activity (29832)
- GO:0004386 : helicase activity (676)
- GO:0003678 : DNA helicase activity (166)

[Get this tree as RDF XML.](#)  
[Get this data as a GO flat file.](#)  
[Get a bookmarkable url of this tree.](#)

In the recent malaria genome paper fatty acid biosynthesis was highlighted as a possible target for chemotherapy.

Using Amigo:

Find the proteins involved in fatty acid biosynthesis in malaria.

Where are many of them they localised within the cell?

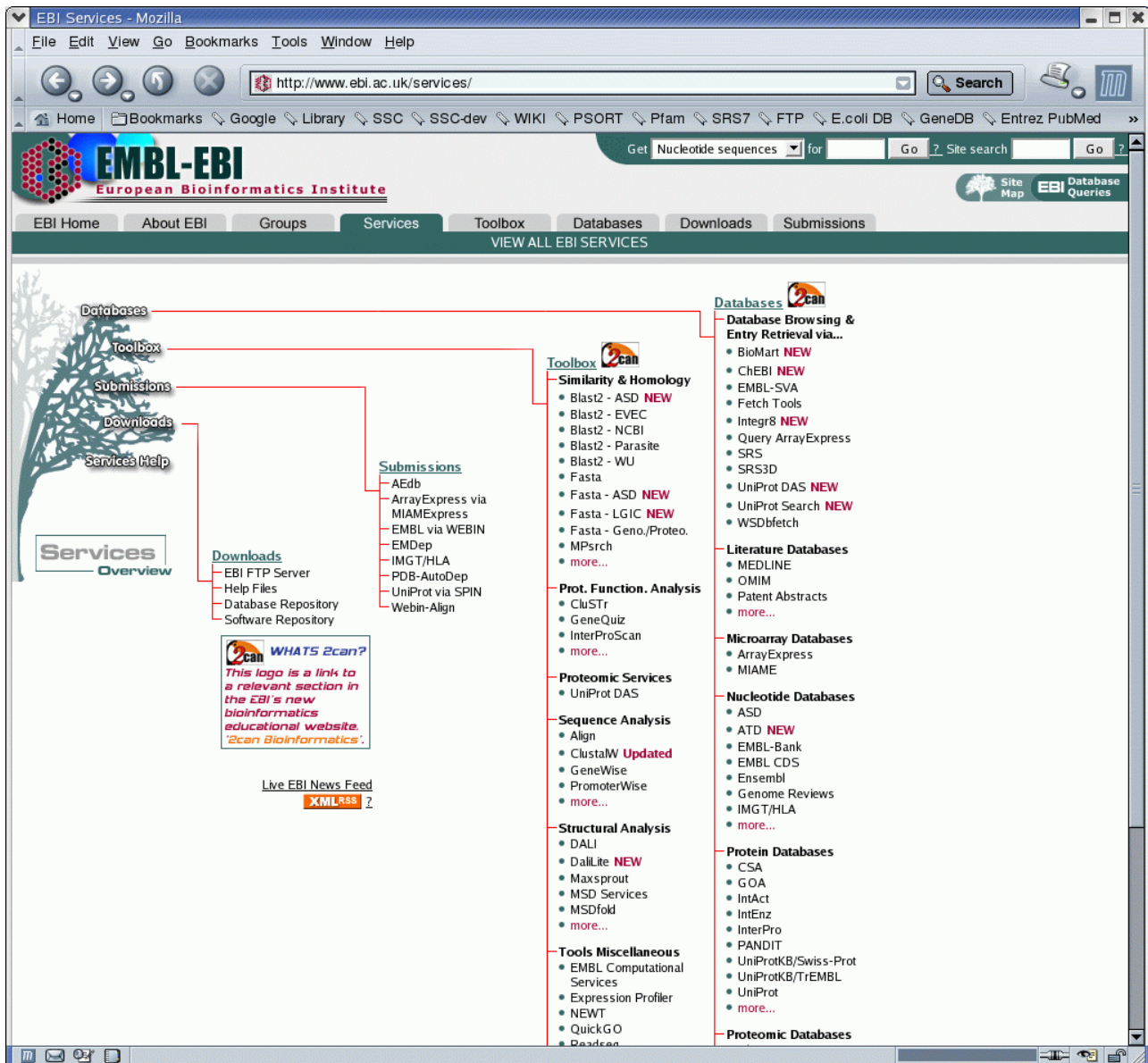
Are there any uncharacterised genes with the same subcellular localisation?

(NB. You can filter using the evidence code to show only those that have had their localisation experimentally confirmed). Ask if you are unclear.

## Section 3

### InterPro & UniProt

Web address: <http://www.ebi.ac.uk/services>



## Exercise 1

Use InteProScan to assign family membership and identify functional domains.

Copy and paste 'Prot1' amino acid sequence into the main box and submit job. You can find the sequence in the "sequences" file under the "Module\_5" directory



InterProScan Sequence Search

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package ([Readme file](#) or [FAQs](#)), or the InterPro [user manual](#) or [help pages](#).

[Download Software](#)

YOUR EMAIL:

RESULTS:

APPLICATIONS TO RUN: ☐ Clear all ☒ Check all

☒ BlastProDom ☒ FPrintScan ☒ HMMPIR ☒ HMMPLam ☒ HMMSmart  
☒ HMMTigr ☒ ProfileScan ☒ ScanRegExp ☒ SuperFamily ☒ SignalPHMM  
☒ TMHMM ☒ HMMPanther

TRANSLATION TABLE (DNA/RNA only):

MIN. OPEN READING FRAME SIZE:

Enter or Paste a PROTEIN Sequence in any format:

Upload a file:

PLEASE NOTE: Interactive job results are stored for 24 hours, email job results are stored for one week.

Paste Prot1  
sequence here

Submit job

InterProScan Results

Table View | Raw Output | XML Output | Original Sequences | SUBMIT ANOTHER JOB

SEQUENCE: [Sequence\\_1](#) CRC64: 91753E248DECA5B1 LENGTH: 548 aa

InterPro <a href="#">IPR001844</a> Family	Chaperonin Cpn60	CHAPERONIN60
InterPro <a href="#">PR00298</a>		CHAPERONINS_CPN60
InterPro <a href="#">PS00296</a>		
InterPro <a href="#">IPR002423</a> Family	Chaperonin Cpn60/TCP-1	TCOMPLEXTCP1
InterPro <a href="#">PR00304</a>		Cpn60/TCP-1
InterPro <a href="#">PTHR11353</a>		Cpn60_TCP1
InterPro <a href="#">PF00118</a>		
InterPro <a href="#">IPR008950</a> Domain	GroEL-like chaperone, ATPase	GroEL-ATPase
InterPro <a href="#">SSF48592</a>		
InterPro <a href="#">IPR012723</a> Family	chaperonin GroEL	GroEL
InterPro <a href="#">TIGR02348</a>		

Table View | Raw Output | XML Output | Original Sequences | SUBMIT ANOTHER JOB

Please contact [EBI Support](#) with any problems or suggestions regarding this site.  
[View Printer-friendly version of this page](#) | [Terms of Use](#)

Click on IPR001844  
to see what  
information you can  
gain about this  
domain

Looking at the window from the page before:

- What domains/sites does this protein contain?
- Click on IPR001844 to see what information you can gain about this domain and look for GO terms that could be assigned to it.
- Scroll down the web page and look for the relationships that this entry might have with other InterPro entries.

InterPro: IPR001844 Chaperonin Cpn60 - Mozilla

http://www.ebi.ac.uk/interpro/Entry?ac=IPR001844

EMBL-EBI European Bioinformatics Institute

EBI Home About EBI Groups Services Toolbox Databases Downloads Submissions

InterPro home Text Search Sequence Search Databases Documentation FTP site Protein of the month

Search: Search Entries Search InterPro

Simple | Full HTML Version Click here for help!

### InterPro IPR001844 Chaperonin Cpn60

**Matches** Overview: sorted by AC, sorted by name, of known structure, grouped by taxonomy, proteins with splice variants  
 Detailed: sorted by AC, sorted by name, of known structure, proteins with splice variants  
 Table: For all matching proteins, of known structure  
[Architectures](#)

**Accession** IPR001844 Chaprin\_Cpn60 Matches: 1526 proteins

**Type** Family

**Signatures**

Database	ID	Name	Proteins
PRINTS	PR00298	CHAPERONIN60	1503
PROSITE pattern	PS00296	CHAPERONINS_CPN60	845

**Children** IPR012723 chaperonin GroEL

**Parent** IPR002423 Chaperonin Cpn60/TCP-1

**Contains** IPR008950 GroEL-like chaperone, ATPase

**Process** GO:0044267 cellular protein metabolism

**Function** GO:0005515 protein binding  
GO:0005524 ATP binding

**Abstract**

The assembly of proteins has been thought to be the sole result of properties inherent in the primary sequence of polypeptides themselves. In some cases, however, structural information from other protein molecules is required for correct folding and subsequent assembly into oligomers [1]. These 'helper' molecules are referred to as molecular chaperones, a subfamily of which are the chaperonins [2]. They are required for normal cell growth (as demonstrated by the fact that no temperature sensitive mutants for the chaperonin genes can be found in the temperature range 20 to 43 degrees centigrade [1]), and are stress-induced, acting to stabilise or protect disassembled polypeptides under heat-shock conditions [2]. Type I chaperonins present in eubacteria, mitochondria and chloroplasts require the concerted action of 2 proteins, chaperonin 60 (cpn60) and chaperonin 10 (cpn10). Type II chaperonins, found in eukaryotic cytosol and in Archaeobacteria, comprise only a cpn60 member.

The 10 kDa chaperonin (cpn10 - or groES in bacteria) exists as a ring-shaped oligomer of between 6 to 8 identical subunits, whereas the 60 kDa chaperonin (cpn60 - or groEL in bacteria) forms a structure comprising 2 stacked rings, each ring containing 7 identical subunits [1]. These ring structures assemble by self-stimulation in the presence of Mg<sup>2+</sup>-ATP. The central cavity of the cylindrical cpn60 tetradecamer provides as isolated environment for protein folding whilst cpn-10 binds to cpn-60 and synchronizes the release of the folded protein in an Mg<sup>2+</sup>-ATP dependent manner [3, 2]. The binding of cpn10 to cpn60 inhibits the weak ATPase activity of cpn60.


GO term



If you look at the top of the page, on the “*Detailed view*” line, follow the link to “*of known structure*” and look at the “*Structural features*” on the first protein (red circle). Note that it has a PDB structure (green stripped bar) for whole length of the protein and two ways of classifying the same domain: CATH (pink stripped bars) and SCOP (black stripped bars). Click on the links to see the differences between these two databases.





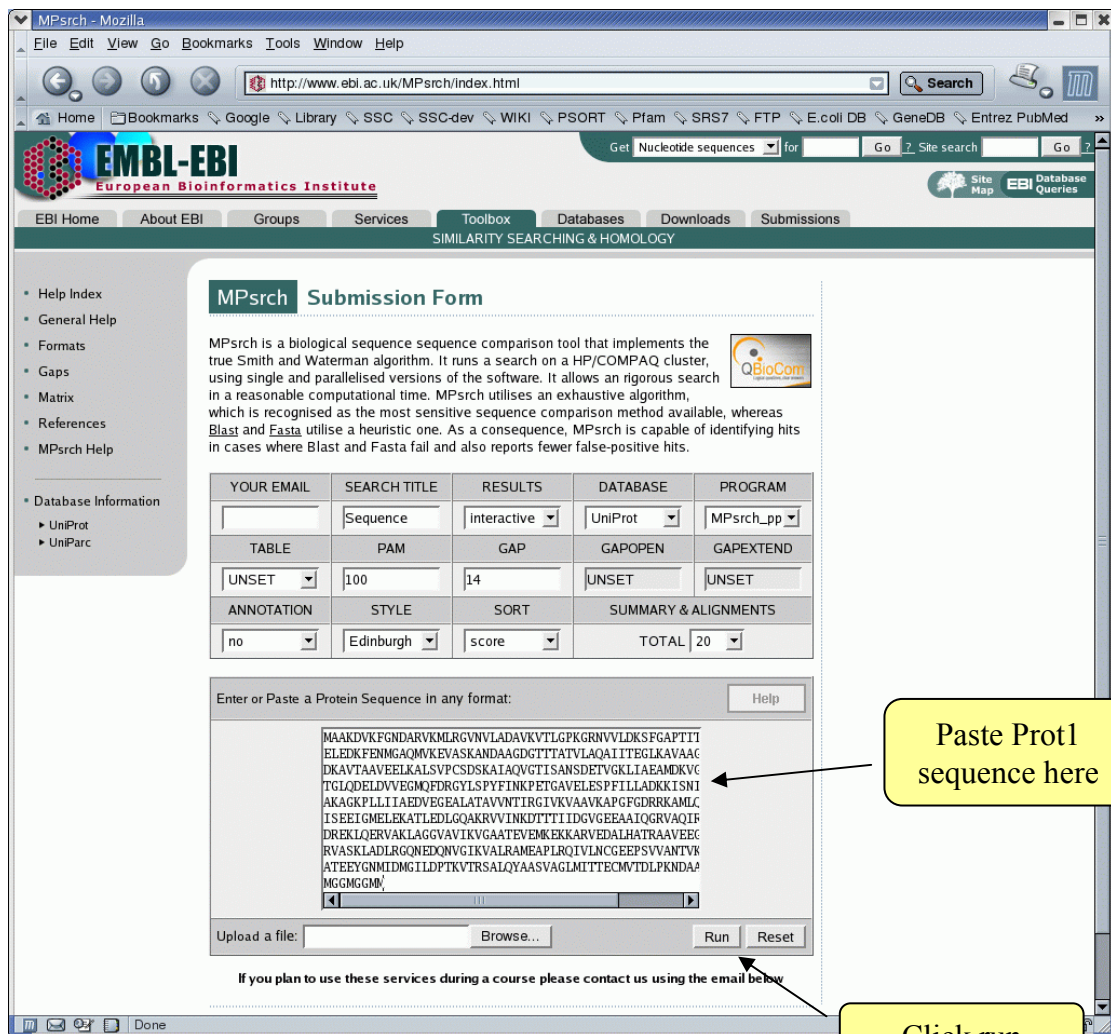
Click on the  symbol adjacent to the CATH domain to have a look at the domain using Astexviewer. Note that the selected CATH domain is highlighted in yellow.

## Exercise 2

Alignments: MPsrch is capable of identifying hits in cases where Blast and Fasta fail.

Web page: <http://www.ebi.ac.uk/MPsrch/index.html>

Copy and paste Prot1 amino acid sequence into the main box and click run. You can find the sequence in the “sequences” file under the “Module\_5” directory



**MPsrch Submission Form**

MPsrch is a biological sequence sequence comparison tool that implements the true Smith and Waterman algorithm. It runs a search on a HP/COMPAQ cluster, using single and parallelised versions of the software. It allows an rigorous search in a reasonable computational time. MPsrch utilises an exhaustive algorithm, which is recognised as the most sensitive sequence comparison method available, whereas Blast and Fasta utilise a heuristic one. As a consequence, MPsrch is capable of identifying hits in cases where Blast and Fasta fail and also reports fewer false-positive hits.

YOUR EMAIL	SEARCH TITLE	RESULTS	DATABASE	PROGRAM
<input type="text"/>	Sequence	interactive	UniProt	MPsrch_pp

TABLE	PAM	GAP	GAOPEN	GAPEXTEND
UNSET	100	14	UNSET	UNSET

ANNOTATION	STYLE	SORT	SUMMARY & ALIGNMENTS
no	Edinburgh	score	TOTAL 20

Enter or Paste a Protein Sequence In any format: Help

```

MAAKDVKFGNDARVKMLRGVNVLDVAVKVTLGPKGRNVVLDKSFAPPTTI
ELEDKFENMGQMVKEVASKANDAAGDGTITATVLAQAIITTEGLKAVAAAC
DKAVTAAVEELKALSVPCSDSKAIAQVGTISANSDETGVKLI AEAMDKVC
TGLQDELVDVVGMDRGLSPYFINKPETGAVELESFILLADKKISNI
AKAGKPLLI AEADVEGEALATAVNTIRGIVKVAVKAPFGDRRKAMLC
TSEETGMELEKATLEDLQAKRVVINKDTITIDGVGEEAATQGRVAQIF
DRKLIQERVAKLAGGVAVIKGAATEVEMKEKKARVEDALHATRAAVEEC
RVASKLADLRGQEDQNVGKVALRAMEAPLRQIVINCGEEPSVVANTVK
ATEEYGNMIDMGILDPKTVRSALQYAASVAGLMTTECMVTDLPKNDAP
MGMGMGM
  
```

Upload a file:  Browse... Run Reset

If you plan to use these services during a course please contact us using the email below

Click on "Show Alignments" to display aligned sequences

**MPsrch Summary Table**

SUBMISSION PARAMETERS			
Title	Sequence	Database	uniprot
Sequence length	548	Sequence type	p
Program	MPsrch_pp	Version	4.2.80
Matrix	PAM 100	Open gap penalty	14
Gap extension penalty	14		

Buttons: Show Annotation, MPsrch Result, XML, SUBMIT ANOTHER JOB, Show Alignments, Clear all, Check all, Invert selection, Reset

Alignment	DB-ID	Description	Length	Match%	Query Match%	Score	Pred.No.
1	<a href="#">UNIPROT:Q548M0_ECOLI</a>	GroEL	548	100.0	100.0	4504	0.00e+00
2	<a href="#">UNIPROT:Q548M1_ECOLI</a>	GroEL	548	99.6	99.6	4486	0.00e+00
3	<a href="#">UNIPROT:Q6UDB1_ECOLI</a>	GroEL (Fragment)	548	99.5	99.5	4483	0.00e+00
4	<a href="#">UNIPROT:Q6Q999_ECOLI</a>	GroEL	548	99.5	99.5	4480	0.00e+00
5	<a href="#">UNIPROT:Q6UDB3_ECOLI</a>	GroEL (Fragment)	548	99.5	99.4	4476	0.00e+00
6	<a href="#">UNIPROT:Q6UDB2_ECOLI</a>	GroEL (Fragment)	548	99.5	99.4	4476	0.00e+00
7	<a href="#">UNIPROT:CH60_ECOLI</a>	60 kDa chaperonin (Protein Cpn60) (groEL protein).	547	99.6	99.3	4473	0.00e+00
8	<a href="#">UNIPROT:CH60_ECO57</a>	60 kDa chaperonin (Protein Cpn60) (groEL protein).	547	99.6	99.3	4473	0.00e+00
9	<a href="#">UNIPROT:CH60_SHIFL</a>	60 kDa chaperonin (Protein Cpn60) (groEL protein).	547	99.6	99.3	4473	0.00e+00
10	<a href="#">UNIPROT:CH60_ECOL6</a>	60 kDa chaperonin (Protein Cpn60) (groEL protein).	547	99.6	99.3	4473	0.00e+00
11	<a href="#">UNIPROT:Q6UDB0_ECOLI</a>	GroEL (Fragment)	548	99.5	99.3	4471	0.00e+00
12	<a href="#">UNIPROT:Q6UDB5_ECOLI</a>	GroEL (Fragment)	548	99.3	99.2	4470	0.00e+00

You can click on the UniProt links for each entry and browse through the different links you find in both the 'Basic' and the 'Extended' web pages.

You can click on the UniProt links for each entry and browse through the different links you find in both the "Basic" and the "Extended" web pages.

**UniProt**  
the universal protein knowledgebase

Basic UniProtKB Entry Viewer

Protein Q548M0\_ECOLI

Buttons: New Query, Submit Annotation, Download Protein, Bookmark Protein (Ctrl-D)

Views: Fasta | Flat File | XML | ExPASy | SRS | PIR

**Basic / Extended**

General information about the UniProtKB/TrEMBL entry

Entry name	Q548M0_ECOLI
Primary accession number	Q548M0
Entered in TrEMBL	Release 31, 13-SEP-2005
Sequence was last modified	Release 31, 13-SEP-2005
Annotations were last modified	Release 31, 13-SEP-2005

**Protein description**

Protein name	GroEL
--------------	-------

**Origin of the protein**

From	Escherichia coli[TaxID:562]
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.

**References**

[1]	NUCLEOTIDE SEQUENCE. STRAIN= JM105; Dugourd D.F., Young A., Wright J.A.; "Molecular chaperones/chaperonin-encoding stress genes groEL and groES and their use as antimicrobial targets."; Submitted (DEC-2000) to the EMBL/GenBank/DBJ databases.
-----	---

**Comments**

FUNCTION	Prevents misfolding and promotes the refolding and proper assembly of unfolded polypeptides generated under stress conditions (By similarity).
SUBUNIT	Oligomer of 14 subunits composed of two stacked rings of 7 subunits (By similarity).

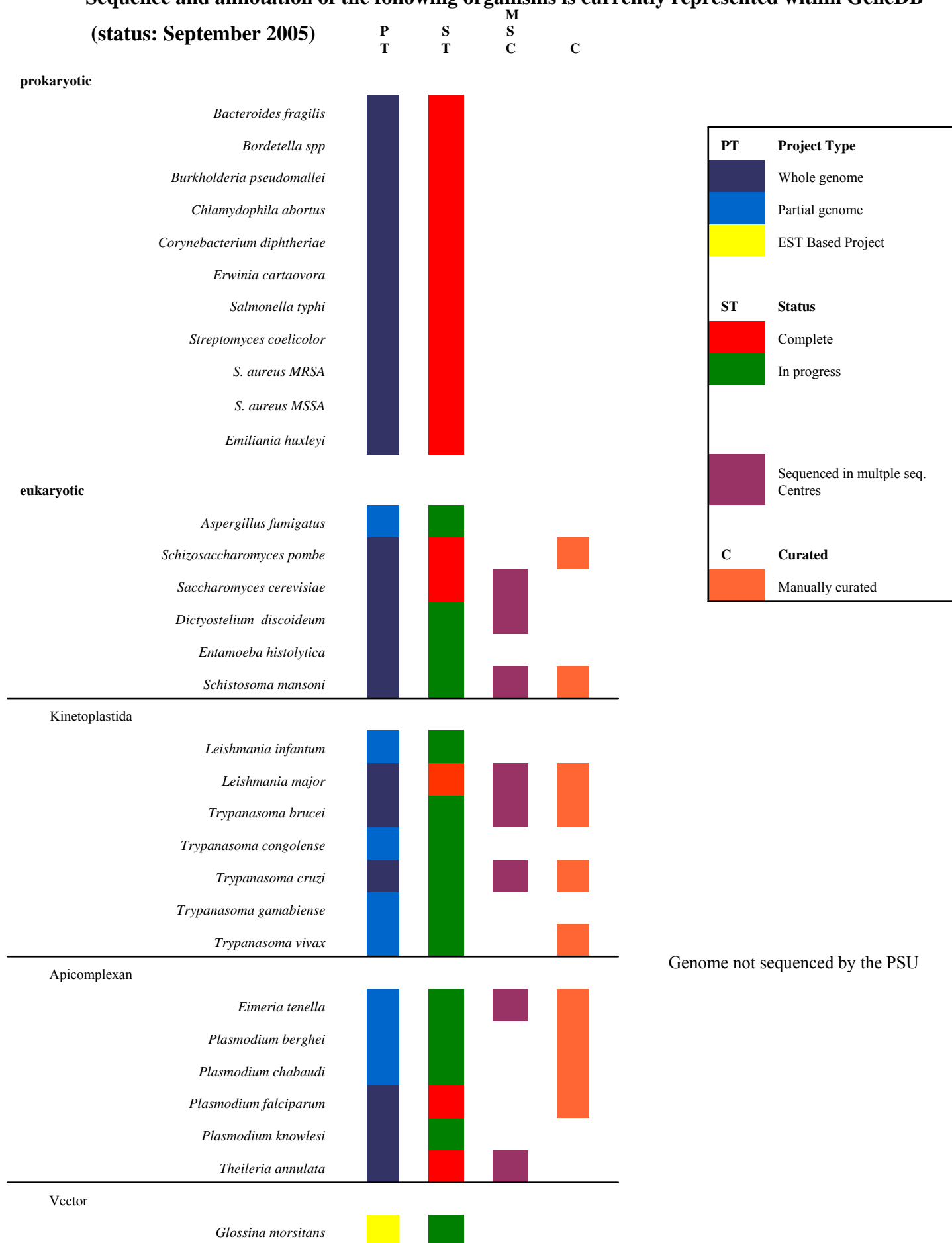
# Module 6

## Data mining using GeneDB

### **Introduction**

This module will demonstrate GeneDB (<http://www.genedb.org>), a genome database housing sequence and annotation of prokaryotic and eukaryotic organisms. The resource provides a portal through which data generated by the Pathogen Sequencing Unit and other collaborating sequencing centres can be made publicly available. It combines data from finished and ongoing genome and expressed sequence tag (EST) projects with curated annotation, that can be searched, sorted and downloaded, using a single web based resource. The current release stores 33 datasets (see Table 1) of which 12 are curated and maintained by biologists, who review and incorporate information from the scientific literature, public databases and the respective research communities.

**Sequence and annotation of the following organisms is currently represented within GeneDB**  
**(status: September 2005)**



## Aims

The aim of this module is for you to familiarise yourself with GeneDB and the various ways of accessing, querying, browsing and retrieving data. You'll use GeneDB as a research tool to retrieve candidate genes which you could follow up with further experimental validation. In the process you will also hopefully see how GeneDB integrates diverse biological datasets, organises, indexes and extensively cross-references these.

In addition, the exercises are designed to make more general points which need to be taken into consideration when approaching and evaluating database searches, not just GeneDB. These are:

1. How complete/incomplete is the dataset you are searching? In the case of organisms with two sets of each chromosome ( i.e. diploid organisms), does the dataset represent the haploid or diploid genome content?
2. How was the dataset generated?
  - a. Is it an EST project? What estimated coverage does the dataset represent (i.e. is it a partial 3-5x coverage or a 8-10x coverage project?)
  - b. Has the sequence been “manually” finished (i.e. sequencing gaps closed and base checked)?
  - c. How were the gene predictions carried out (automated vs. manual)?
  - d. How were the gene prediction annotated (automated vs. manual)?
3. Depending on the gene prediction and associated annotation method, you may need to approach querying from several angles, not just 1 methodology (e.g. combine keyword searching with similarity searching).
4. When designing your searches:
  - a. identify keywords that describe your topic.
  - b. identify any synonyms for your keywords.
  - c. be aware of spelling variations and/or plurals.
  - d. decide the scope of your search.
  - e. be aware that using the same search method in different databases may affect your results.
  - f. try different search methods to identify candidate genes.
  - g. be aware of the use of wildcards.

### **Exercise 1:**

Data mining the *T. brucei* genome for the Arp2/3 complex.

### **Exercise 2:**

Using the Artemis applet to retrieve sequence and annotated features.

### **Exercise 3:**

Demonstration of the Boolean querying tool.

### **Exercise 4:**

Data mining of *Plasmodium* genomes for monosaccharide transporters.

### **Exercise 5:**

Data mining three *Bordetella* genomes for autotransporter genes



## Exercise 1 Data mining of the *T. brucei* genome for the Arp2/3 complex.

Can you identify the components of the Arp2/3 complex in the kinetoplastid organism *Trypanosoma brucei*, causative agent of sleeping sickness in sub-Saharan Africa?

The Arp2/3 complex is involved in actin assembly and function in the eukaryotic cytoskeleton. So far this complex has not been investigated in kinetoplastids, but has been well characterised in other organisms, such as the fission yeast *Schizosaccharomyces pombe*. Unlike the *S. pombe* genome, which is complete and contains extensive curated annotations, the genomes of the trypanosomatids are in various stages of completion and annotation. Using a variety of tools/methods, some of which you will already have covered in earlier modules, identify putative members of this complex and complete the table on page at the end of this exercise.

Start by identifying how many components have been annotated to this complex in *S. pombe*, which you will be using as a thoroughly annotated reference genome (**exercise 1.1**).

### Exercise 1.1

1

Go to the GeneDB homepage (<http://www.genedb.org>)



Welcome to the  
GeneDB website  
Version 2.1



Database Entry Point

Searches	Sequence Searches	Datasets
Search for gene by ID/description in All organisms <input type="text"/> <input checked="" type="checkbox"/> Include description in search <input type="checkbox"/> Add wildcards to search term <input type="button" value="Reset"/>	<a href="#">omniBLAST</a> (Multi-organism BLAST) <input type="button" value="Go To"/> single organism BLAST: <input type="text"/>	Fungi <input type="button" value="Go To"/> Choose... <input type="button" value="Go To"/> Choose... A. fumigatus S. cerevisiae S. pombe Bacteria <input type="button" value="Go To"/> Choose... Parasite Vectors <input type="button" value="Go To"/> Choose...

2

Select *S. pombe* from  
the pull down menu

Go to our [main search page](#), our [complex querying page](#) or [AmiGO](#)

Information	Links
<a href="#">Guide to GeneDB</a> What is GeneDB, and what's in it? Navigating/Searching GeneDB Contacting Us/Feedback Privacy Policy <a href="#">Data Release Policy</a>	PSU Sequencing Projects Software <a href="#">Prokaryotes</a> <a href="#">Eukaryotes (Protozoa)</a> <a href="#">Eukaryotes (Fungi)</a> <a href="#">ACT</a> <a href="#">Artemis</a>

Funded as part of the Wellcome Trust Functional Genomics Development Initiative, the GeneDB project has a primary goal: to develop and maintain curated database resources for three organisms: *Schizosaccharomyces pombe*, for which a complete genome sequence has been obtained, and the kinetoplastid protozoa *Leishmania major* and *Trypanosoma brucei*, whose genome sequences have yet to be completed. It is envisaged that the generic database structure will subsequently be adopted to integrate datasets for other organisms, both prokaryotic and eukaryotic, that have been sequenced by the Sanger Institute Pathogen Sequencing Unit.

**Note:** as this site is under [continual development](#), please be patient if things do not appear to be working - let us know, and it will be fixed. Please also note that the URLs (web page addresses) may change, so [be cautious about creating links to them](#). If you contact the developers we will try and keep you informed of any changes. If you have any suggestions about the site please [contact the developers](#)



3

Type **Arp2/3** into the search box, ensuring that the 'Add wildcards to search term' box is also selected. This will search all the gene names and product/description lines assigned to features within the *S. pombe* dataset. Not selecting the wildcard box would only return exact matches.

**GeneDB** *Schizosaccharomyces pombe* GeneDB **GeneDB**

WARNING: 20th July 2005. There is currently a bug with the downloading of intergenic sequences from the List/Report Download if partial CDS are included. This feature will work correctly if the partial CDS sequences SPAC977.01 SPAC750.08C or SPAPB21E7.10 are not included in your download list.

This database contains all *S. pombe* known and predicted protein coding genes, pseudogenes, transposons, tRNAs, rRNAs, snRNAs, snoRNAs and other known and predicted non-coding RNAs. Curation of new and existing literature is ongoing and changes are incorporated weekly.

Database Entry Point Go To Organisms

Search for  by ID/description  
  
☒ Include description  
☒ Add wildcards

Searches/Analysis  
[omniBLAST](#)  
[BLAST](#)  
[Motif Search](#)  
[EMOWSE](#)  
[AmiGO](#)  
[List Download](#)  
[Cross-Organism Search Page](#)  
[Complex/Boolean Query](#)

Browse Catalogues  
[Products](#)  
[Curation](#)  
[SWISS-PROT Keywords](#)  
[Pfam](#)  
[Genome Browser](#)  
[Contig/Chromosome Maps](#)

Information  
[Gene Name Registry](#)  
[Project Information](#)  
[Other useful websites](#)  
[Download data \(by FTP\) using "Download Datasets" topics](#)  
[Subscribe to the pombe mailing list](#)  
[S. pombe Project Page](#)

Genome News  
**July 2005**  
**Second East Coast Regional pombe Meeting**  
 This meeting will take place from November 11-13, 2005 in Miami Beach, Florida. Mail [info@USpombe2005.org](mailto:info@USpombe2005.org) for general enquiries and see the [meeting website](#) for details.  
**February 2005**  
**European Fission Yeast Meeting**  
 An interim *S. pombe* meeting will take place on 16th-18th March 2006 at the Wellcome Trust Genome Campus in Hinxton (Cambridge, UK). Preliminary details are available at the [Wellcome Trust Conference Programme website](#).  
[Previous news items](#)

Example Genes  
[Shortcuts to frequently requested gene lists](#)  
 Characterised genes [cdc2 rad1 dld1](#)  
 Inferred function [SPAC16E8.04c](#) [SPAC26H5.02c](#)  
 Conserved hypothetical [SPCC830.08c](#) [SPCC4B3.13](#)  
 Sequence orphans [SPCC594.07c](#) [SPCC70.10](#)

4

The results list below will be returned. Clicking on the hyperlinked gene names will take you to the feature page. Click on SPAC6G9.07c to see the information contained on a feature page.

**GeneDB** **Gene Results List**

Go To Organisms Go To Shortcuts [Help](#)

Results 1 to 6 of 6 results shown

[Previous](#) [Next](#) [Report Download](#)

<i>S. pombe</i>	CDS	<a href="#">SPAC6G6.10c</a>	arc2, arc2, ARP2/3 actin-organizing complex subunit Arc34, arc34
<i>S. pombe</i>	CDS	<a href="#">SPAC6G9.07c</a>	arc4, ARP2/3 actin-organizing complex subunit Arc4, obsolete arp10
<i>S. pombe</i>	CDS	<a href="#">SPAC11H11.06</a>	arp2, ARP2/3 actin-organizing complex subunit Arc2, SPAC22F8.01
<i>S. pombe</i>	CDS	<a href="#">SPBC14C8.06</a>	arc1, ARP2/3 actin-organizing complex subunit Sop2, sop2
<i>S. pombe</i>	CDS	<a href="#">SPAC17G8.04c</a>	arc5, ARP2/3 actin-organizing complex subunit Arc16, arc16
<i>S. pombe</i>	CDS	<a href="#">SPBC1778.08c</a>	arc3, arc3, ARP2/3 actin-organizing complex subunit Arc21, arc21

[Previous](#) [Next](#)

Hosted by the [Sanger Institute](#) [Send comments, requests, corrections and updates](#)

Click on the links on the feature page to see how the data are cross-linked and referenced.

**GeneDB** CDS: arc4

WARNING: 20th July 2005. There is currently a bug with the downloading of intergenic sequences from the List/Report Download if partial CDS are included. This feature will work correctly if the partial CDS sequences SPAC77.01 SPAC750.082 or SPAPB21E7.10 are not included in your download list.

Search for:  Go To: Organisms  Go To: Shortcuts  [Help](#) [Contact curator](#)

Biological process, [cellular](#) component and molecular function annotation is being removed from the description part gene pages and replaced by manual GO curation. To obtain the full complement of gene products annotated to specific terms please use the links to Amigo in the left hand column of the "Gene Ontology" section of the page.

**General Information** [Add to Basket](#) [View Basket](#)

Name: arc4  
Systematic Name: SPAC6G9.07c  
Obsolete Synonyms: arp10  
Status: role inferred from homology  
Product: ARP2/3 actin-organizing complex subunit Arc4  
Type: CDS  
Sequence: [DNA and Protein](#)

**Location**

Chromosome: 1  
Contig: c977  
Contig Location: complement(3228809..3229681) (Unspliced length: 873 bp)  
Exons: complement(join(3228809..3228983), 3229167..3229290, 3229421..3229623, 3229679..3229681)) (Spliced length: 507 bp)

[Graphical Display \(in Artemis\)](#) [Genome Browser](#)

Context Map:

SPAC6G9.03c SPAC6G9.04 SPAC6G9.05 pcpl >arc4< ubp6 rpl24 sen1 svb1 cfr1

**Curation**

ARP2/3 actin-organizing complex subunit Arc4  
similar to *S. cerevisiae* [YKL019C](#)

**Predicted Peptide Properties**

Mass	19.6 kDa	Amino acids	168
Isoelectric point	pH 5.2	Charge	-3.0

Signal Peptide: Not found  
Transmembrane Domains: 0 found  
GPI Anchor: Not found

**Gene Ontology Annotation**

Term (browse Amigo)	Qualifier	Evidence	Other genes annotated to this term
<b>Biological Process</b>			
<a href="#">actin cortical patch assembly</a>	ISS (GOC:unpublished)	with <a href="#">SPTR.P33204</a>	<a href="#">15 others</a>
<a href="#">regulation of actin filament polymerization</a>	ISS (GOC:unpublished)	with <a href="#">SPTR.P33204</a>	<a href="#">4 others</a>
<b>Cellular Component</b>			
<a href="#">actin cortical patch</a>	ISS (GOC:unpublished)	with <a href="#">SPTR.P33204</a>	<a href="#">41 others</a>
<a href="#">Arp2/3 protein complex</a>	ISS (GOC:unpublished)	with <a href="#">SPTR.P33204</a>	<a href="#">6 others</a>

**Published Expression Profiles**

Gene Expression Viewer [Cell Cycle](#) [Meiosis](#) [Environmental Stress](#)

**Phenotype**

[FYSSION \(\*S. pombe\* strain database\)](#)

**Literature**

Search for arc4 in [PubMed](#)

**Domain Information**

[View Pfam domain structure for this gene product](#)  
[View SCOP superfamily](#)

DB	Access	Description
Pfam	<a href="#">PF05856</a>	ARP2/3 complex 20 kDa subunit (ARPC4)
InterPro	<a href="#">IPR008384</a>	ARP23 complex 20 kDa subunit

**Database Cross-References**

DB	Access	Description
UniProt	<a href="#">Q92352</a>	Probable ARP2/3 complex 20 kDa subunit (p20-ARC)
EMBL	<a href="#">Z81317</a>	<i>S. pombe</i> chromosome I cosmid c6G9.
EMBL	<a href="#">AB010050</a>	<i>Schizosaccharomyces pombe</i> mRNA for 20 kd actin related protein complex, partial cds.
GermOnline	<a href="#">SPAC6G9.07c</a>	GermOnline
PIR	<a href="#">T39069</a>	PIR
PIR	<a href="#">T43309</a>	PIR
PombePD	<a href="#">SPAC6G9.07c</a>	Proteome, Inc

**SWISS-PROT Annotation For This Protein**

Similarity: Belongs to the ARPC4 family.  
Function: Part of a complex implicated in the control of actin polymerization in cells (By similarity).  
Sub Unit: Belongs to a complex composed of ARP2, ARP3, P41-ARC, P34-ARC, P21-ARC, P20-ARC and P16-ARC (By similarity).  
Keywords: Complete proteome ([4969 others](#)), Cytoskeleton ([25 others](#))

This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.sib-swiss.ch/announcement/> or send an email to [licences@sib-swiss.ch](mailto:licences@sib-swiss.ch)).

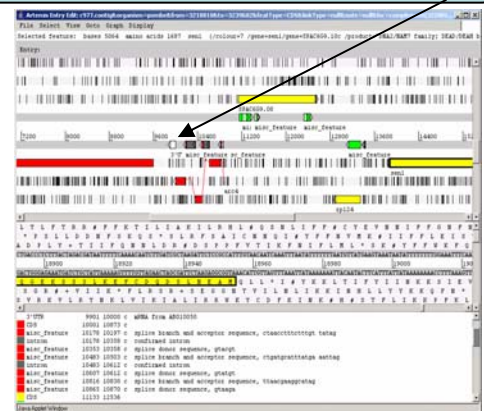
### Navigation bar pull down menus:

You can navigate between different organism datasets and search tools using pull down menus

**Gene name and product information:** The product lines are standardized and indexed so that features sharing the same product lines can be retrieved. Access to the nucleotide and amino acid sequences of the feature are also provided.

### Basic location information and context map:

Clicking on the 'Graphical display in Artemis' open up an Artemis applet – which will be discussed further in exercise 2. Via the applet, the feature can be viewed in the context of the sequence and additional annotation, such as UTRs.




**GO associations:** Links will take you to the descriptions of the terms as well as other proteins annotated to the same ontology node.


**Database cross-references** to literature, phenotype, protein motif/domain as well as sequence databases.

6

Go back to your initial results page and see how many proteins have been assigned to the Arp2/3 complex in *S. pombe*. Are any of the components characterised by Pfam domains? Fill in the table on page at the end of the exercise.



## Gene Results List



Go To Organisms
Go To Shortcuts
[Help](#)

Results 1 to 6 of 6 results shown

Previous
Next

[Report Download](#)

<i>S. pombe</i>	CDS	<a href="#">SPAC6F6.10c</a>	arc2, arc2, ARP2/3 actin-organizing complex subunit Arc34, arc34
<i>S. pombe</i>	CDS	<a href="#">SPAC6G9.07c</a>	arc4, ARP2/3 actin-organizing complex subunit Arc4, obsolete:arp10
<i>S. pombe</i>	CDS	<a href="#">SPAC11H11.06</a>	arp2, ARP2/3 actin-organizing complex subunit Arp2, SPAC22F8.01
<i>S. pombe</i>	CDS	<a href="#">SPBC14C8.06</a>	arc1, ARP2/3 actin-organizing complex subunit Sop2, sop2
<i>S. pombe</i>	CDS	<a href="#">SPAC17G8.04c</a>	arc5, ARP2/3 actin-organizing complex subunit Arc16, arc16
<i>S. pombe</i>	CDS	<a href="#">SPBC1778.08c</a>	arc3, arc3, ARP2/3 actin-organizing complex subunit Arc21, arc21

Previous
Next

Hosted by the [Sanger Institute](#)
[Send comments, requests, corrections and updates](#)

The next aim is to identify putative Arp2/3 complex components in the genomes of trypanosomatids. We're going to start off with the *T. brucei* genome. You may take a number of different approaches:

- using keywords to search the available *T. brucei* annotation (**exercise 1.2**).
- browsing predicted Pfam domain catalogue (**exercise 1.3**).
- using GO annotations and the gene ontology browser (**exercise 1.4**).
- using BLAST to identify sequences with similarity to known Arp2/3 components (**exercise 1.5**).

## Exercise 1.2 The use of keywords to search the available *T. brucei* annotation

1

Go to the GeneDB homepage (<http://www.genedb.org>)

2

Select *T. brucei* as an organism

3

Type Arp2/3 into the search box

The homepages also provide up-to-date information about sequencing progress, data updates, nomenclature and other community resources.



4

The results list below will be returned. Which components of the Arp2/3 complex have already been annotated in the *T. brucei* genome? Have a look by clicking on the hyperlinked gene names. Would you agree with the assignments?

Previous		Next	
		<a href="#">Report Download</a>	
<i>T. brucei</i>	CDS	<a href="#">Tb10.70.2680</a>	ARP2/3 complex subunit, putative, ARP2/3 complex subunit, putative, TRYP_x-70a06.p2kb545_360
<i>T. brucei</i>	CDS	<a href="#">Tb927.8.4410</a>	ARP2/3 complex subunit, putative, Tb08.29H22.800
<i>T. brucei</i>	CDS	<a href="#">Tb10.389.0270</a>	actin related protein 2/3 complex, putative, ARP2/3 complex subunit, putative
<i>T. brucei</i>	CDS	<a href="#">Tb10.406.0320</a>	ARP2/3 complex 16kDa subunit, putative, ARP2/3 complex subunit, putative
<i>T. brucei</i>	CDS	<a href="#">Tb927.2.2900</a>	ARP2/3 complex subunit, putative, 10C8.250

### Exercise 1.3 Browsing the Pfam domain catalogue

As you will have seen from the *S. pombe* example earlier on, some of the subunits are characterised by Pfam domains and you should have made a note of these in the table. Therefore, one way of identifying the putative components would be by browsing the Pfam catalogue.

If you haven't made a note of the Pfam domains, then you could either go back to the *S. pombe* dataset using the navigation bar alternatively, use the Pfam site at <http://www.sanger.ac.uk/Software/Pfam/> to retrieve the domain information by typing **Arp2/3** into the search box.

Family	Description
<a href="#">P34-Arc</a>	Arp2/3 complex, 34 kD subunit p34-Arc
<a href="#">P21-Arc</a>	P21-ARC (ARP2/3 complex: 21 kDa subunit)
<a href="#">ARP2/3</a>	ARP2/3 complex 20 kDa subunit (ARP2/3)
<a href="#">P16-Arc</a>	ARP2/3 complex 16 kDa subunit (p16-Arc)
<a href="#">AARP2CN</a>	AARP2CN (NUC121) domain
<a href="#">Tropomodulin</a>	Tropomodulin
<a href="#">Actin</a>	Actin

**WARNING: 5th September 2005.** The Full Content Search is still pointing at the previous data run's data. This will be fixed later in the day. GeneDB contains release 4 of the *T. brucei* genome (strain TREU927/4 GUTat10.1) generated by the *T. brucei* projects at The Institute for Genomic Research (TIGR's *T. brucei* project) and The Wellcome Trust Sanger Institute (Sanger's *T. brucei* project). It also contains the sequence and annotation of 3 *T. brucei* strain 427, variant 221a bloodstream expression sites ([PMID](#)). Click [here](#) for more information on the *T. brucei* genome/proteome and [here](#) to find out more about the individual chromosome assemblies, in particular with regards to additional unordered contigs from chrIX - XI as well as 3 BACs from homologous regions of chrV, VII and VIII).

1

Select the Pfam link on the *T. brucei* homepage

Database Entry Point

Search for  gene  
by ID/description  
 Arp2/3  
☒ Include description  
☒ Add wildcards

Full Content Search

Searches/Analysis  
[omniBLAST](#)  
[BLAST](#)  
[Motif Search](#)  
[EMOWSE](#)  
[AmiGO](#)  
[List Download](#)  
[Cross-Organism Search Page](#)  
[Complex Boolean Query](#)  
[RNAit \(primer design\)](#)

Browse Catalogues  
[Products](#)  
[SWISS-PROT Keywords](#)  
[Pfam](#)  
[InterPro](#)  
[Genome Browser](#)  
[Contig/Chromosome Maps](#)

Information  
Data  
Example genes  
Data releases  
FTP download (Sanger)  
FTP download (TIGR)  
Help  
Feedback: Curator, Technical

Links  
[Sanger \*T. brucei\* project](#)  
[T. congolense project](#)  
[T. b. gambiense project](#)  
[T. brucei Genome Network](#)  
[TIGR \*T. brucei\* project](#)  
[T. vivax project](#)  
[Biological resources](#)

News  
[News archive](#)  
Trypanosomatids: genomes and biology CD  
15th July 2005  
Tbrnyp genomes published in Science  
22nd June 2005  
non-coding RNA annotation revised  
protein feature/domain predictions updates  
systematic identifiers for chr 3 - 8 assigned by TIGR  
2nd February 2005  
updated *T. vivax* data in GeneDB  
*T. congolense* and *T. b. gambiense* data available via GeneDB

2

The Pfam domain descriptions will be listed alphabetically

Go To:  GeneDB

Results 1 to 100 of 983 results shown

Key: Warning: Automatic prediction

Previous  This list takes you to the first entry that starts with the selected letter, and shows you the next 100 alphabetical entries

0 1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Gold-thioester DNA-binding domain PF00412 (2)

14-3-3 protein PF00944 (1)

2-oxoglutarate-dependent acyltransferase (catalytic domain) PF00188 (2)

2-oxoglutarate-dependent acyltransferase regulatory subunit N-terminal domain PF04653 (1)

2Fe-2S iron-sulfur cluster binding domain PF00111 (4)

2OG-Fe(II) oxygenase superfamily PF03171 (10)

T exonuclease family, domain 1 PF01138 (3)

T-5' exonuclease PF01612 (2)

T5'-cytosine nucleotide phosphodiesterase PF00233 (4)

2-beta-hydroxyacyl-CoA dehydrogenase/iron-sulfur domain PF01073 (1)

3-hydroxyacyl-CoA dehydrogenase, C-terminal domain PF00725 (1)

3-hydroxyacyl-CoA dehydrogenase, NAD binding domain PF02737 (1)

3-oxo-5-alpha-steroid 4-dehydrogenase PF02544 (1)

4F5 protein family PF04419 (1)

4Fe-4S binding domain PF00037 (1)

5'-3' exonuclease, C-terminal SAM fold PF01367 (1)

6-phosphofructo-2-kinase PF01591 (2)

6-phosphogluconate dehydrogenase, C-terminal domain PF00393 (1)

ADP 3-hydroxyacyl-CoA dehydrogenase PF00062 (1)

3

Click on the letter 'A' and scroll down the list

[Arginyl tRNA synthetase N terminal domain PF03485 \(1\)](#)

[Armadillo/beta-catenin-like repeat PF00514 \(5\)](#)

[ARP2/3 complex 16 kDa subunit \(p16-Arc\) PF04699 \(1\)](#)

[ARP2/3 complex 20 kDa subunit \(ARPC4\) PF05856 \(1\)](#)

[Arp2/3 complex, 34 kDa subunit p34-Arc PF04045 \(1\)](#)

[Aspartate/ornithine carbamoyltransferase, Asp/Om binding domain PF00185 \(1\)](#)

[Aspartate/ornithine carbamoyltransferase, carbamoyl-P binding domain PF02729 \(1\)](#)

4

Click on the letter 'P'

Key: Warning: Automatic prediction

Previous  This list takes you to the first entry that starts with the selected letter, and shows you the next 100 alphabetical entries

0 1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

P21-ARC (ARP2/3 complex 21 kDa subunit) PF04062 (1)

PA domain PF02225 (1)

PAF2 superfamily PF01569 (5)

Pagan family cysteine protease PF00112 (12)

Parafagellar rod protein PF05149 (11)

How many of the putative Arp2/3 complex components can be identified using the Pfam catalogue?



### Exercise 1.4 Using GO annotations

By now, you should have identified putative homologues of 5 out of the 7 Arp2/3 complex components. You should be missing the 2 actin related proteins Arp2 and Arp3. Again, there are a number of ways to retrieve possible candidates:

- You could browse the Pfam catalogue for **Actin (PF00022)** – this will give you a short list of 9 candidates as a number of actin related proteins as well as actin itself which share this domain.
- We're going to use GO and similarity searching to identify those last two remaining components.

Gene Ontologies are structured vocabularies that are designed to describe biological processes in an accurate and consistent way (for more information see <http://www.geneontology.org>.) It is composed of three separate ontologies, describing aspects of a given protein's function in terms of its molecular function, biological process and cellular component (location). Where evidence exists from the literature, from sequence analysis or other sources gene ontology terms for function, process and component are attributed to that gene. AmiGO is the database housing assigned gene ontology associations and is maintained by the Gene Ontology consortium. It allows searching and browsing of gene ontology annotations across many genomes from human, mouse through to lower eukaryotes (including those which are not annotated and curated for GeneDB). GeneDB has a copy of the GO database and an installation of the AmiGO browser on top of it. Advantages of a local copy of the GO database include an increased update frequency as well as the inclusion of datasets not otherwise searchable via the 'official' GO database (e.g. assignments inferred by electronic annotation).

It can be a powerful way to search for genes with similar function across several organisms. The example below shows how to set up this query, which can be either accessed from the organism home page and/or the search menu bar at the top of each of the feature pages. Once you've tried it and have become familiar with it, try some of the other suggested searches or perhaps one that would be of interest to your own research.

1

Go to the feature page for Tb927.2.2900, one of the candidates you've come across earlier. You can type the gene name straight into the search box at the top of a feature page.

**GeneDB** CDS: Tb927.2.2900 **GeneDB**

WARNING: 5th September 2005. The Full Content Search is still pointing at the previous data run's data. This will be fixed later in the day.

Search for:  Go To:  Organisms:  Go To:  Shortcuts:  [Help](#) [Contact curator](#)

The sequence and annotation on this page were provided by [TIGR's T. brucei project](#).  
This gene is in [TIGR's T. brucei Annotation Database \(TGAD\)](#).

**General Information** [Add to Basket](#) [View Basket](#)

Name: Tb927.2.2900  
Systematic Name: Tb927.2.2900  
Gene Synonyms: 10CB 250  
Status: role inferred from homology  
Products: ARP2/3 complex subunit, putative (5 others)  
Type: CDS  
Sequence: [DNA and Protein](#)

**Location**

Chromosome: 2  
Chromosome Location: 554106..554648 Length: 543 bp  
[Graphical Display \(in Artemis\)](#) [Genome Browser](#)

Context Map:

[Tb927.2.2830](#) [Tb927.2.2860](#) [Tb927.2.2880](#) [Tb927.2.2900](#) [Tb927.2.2910](#) [Tb927.2.2920](#) [Tb927.2.2940](#) [Tb927.2.2950](#) [Tb927.2.2970](#)

**Primary Annotation**

ARP2/3 complex subunit 4 (curated by B. Wickstead, Univ. of Oxford)

**Predicted Peptide Properties**

Mass	20.7 kDa	Amino acids	180
Isoelectric point	pH 7.9	Charge	3.0

Signal Peptide: Not found  
Transmembrane Domains: 0 found  
GPI Anchor: Not found

Protein Map:

**Domain Information**

DB	Accession	Description	Note
Pfam	PF00835	ARP2/3 complex 20 kDa subunit (ARP2/3)	Residue: 1-177 (Score: 1.8e-60)
InterPro	IPR008384	ARP2/3 complex 20 kDa subunit	Derived from hit Pfam

**Gene Ontology Annotation**

Term (reverse Amigo)	Qualifier	Evidence	Other genes annotated in this term
<a href="#">actin filament polymerization</a>	ISS (TIGR_Tb927.2.2900(TIGR_REF_GO_ref) with SWISS-PROT O15509	1 other	
<a href="#">actin filament polymerization</a>	IEA (GOCCentrepro2go)	1 other	
<a href="#">Cellular Component</a>			
<a href="#">Arp2/3 protein complex</a>	ISS (TIGR_Tb927.2.2900(TIGR_REF_GO_ref) with SWISS-PROT O15509	2 others	
<a href="#">cytoskeleton</a>	IEA (GOCCentrepro2go)	126 others	
<a href="#">Molecular Function</a>			
<a href="#">structural constituent of cytoskeleton</a>	ISS (TIGR_Tb927.2.2900(TIGR_REF_GO_ref) with SWISS-PROT O15509	20 others	

Search for gene/protein in [PubMed](#)

**Database Cross-References**

DB	Accession	Description
UniProt	<a href="#">Q7YVE2</a>	Actin-related protein, ARP2/3 complex subunit, putative
TGAD	<a href="#">Tb927.2.2900</a>	TIGR's T. brucei Annotation Database (TGAD)

**Orthologues**

DB	Accession	Description	Note
GeneDB_Lmajeor	<a href="#">LmJF02.0600</a>	ARP2/3 complex subunit, putative	predicted by jaccard cog clustering
GeneDB_Terui	<a href="#">Tb00.1047053509127.104</a>	ARP2/3 complex subunit, putative	predicted by jaccard cog clustering
GeneDB_Terui	<a href="#">Tb00.1047053508737.194</a>	ARP2/3 complex subunit, putative	predicted by jaccard cog clustering

**Database Similarities**

2

Arp2 can be classified as a structural constituent of the cytoskeleton. Click on the link to other genes annotated to the GO term to see which other 20 proteins have been annotated to this GO term.

3

Can you identify the two missing actin-related proteins Arp2 and Arp3 from this list?

All 19 results shown

[Export Download](#)

T. brucei	CDS	<a href="#">Tb07.28812.620</a>	I/6 autoantigen	
T. brucei	CDS	<a href="#">Tb07.28812.630</a>	I/6 autoantigen	
T. brucei	CDS	<a href="#">Tb09.160.3850</a>	actin-related protein 3, putative, 28G16.235, chr9contg160 tmp0298, obsolete chr9 tmp. 121c, obsolete chr9_contg3519 tmp. 166c	★
T. brucei	CDS	<a href="#">Tb09.160.3960</a>	actin-like protein, putative, 28G16.280, chr9contg160 tmp0305, obsolete chr9 tmp. 116c, obsolete chr9_contg3519 tmp. 159c	★
T. brucei	CDS	<a href="#">Tb09.211.0620</a>	actin, chr9contg211 tmp0078c	
T. brucei	CDS	<a href="#">Tb09.211.0630</a>	actin, chr9contg211 tmp0079c	
T. brucei	CDS	<a href="#">Tb10.61.0500</a>	actin-related protein 2, putative, TRYF_x-61b03 q1kb185_78	
T. brucei	CDS	<a href="#">Tb10.70.5830</a>	actin-like protein, putative, TRYF_x-70a06 p1kb545_795	
T. brucei	CDS	<a href="#">Tb11.01.1870</a>	actin-like protein, putative	
T. brucei	CDS	<a href="#">Tb11.02.1380</a>	TRYFARP, actin-like protein, putative	
T. brucei	CDS	<a href="#">Tb927.1.2330</a>	beta tubulin	
T. brucei	CDS	<a href="#">Tb927.1.2340</a>	alpha tubulin	
T. brucei	CDS	<a href="#">Tb927.1.2350</a>	beta tubulin	
T. brucei	CDS	<a href="#">Tb927.1.2360</a>	alpha tubulin	
T. brucei	CDS	<a href="#">Tb927.1.2370</a>	beta tubulin	
T. brucei	CDS	<a href="#">Tb927.1.2380</a>	alpha tubulin	
T. brucei	CDS	<a href="#">Tb927.1.2390</a>	beta tubulin	
T. brucei	CDS	<a href="#">Tb927.1.2400</a>	alpha tubulin	
T. brucei	CDS	<a href="#">Tb927.2.2900</a>	ARP2/3 complex subunit, putative, 10CB 250	

## Exercise 1.4 Exploring the GeneDB AmiGO browser further

- 1 The local AmiGO browser not only provides access to *T. brucei* terms annotated to them same but also to proteins from other organisms annotated to this term. To see which other eukaryotic proteins have been annotated to this term, click on the hyperlinked term 'structural constituent of the cytoskeleton'.

The sequence and annotation on this page were provided by [TIGR's T. brucei project](#).  
This gene is in [TIGR's T. brucei Annotation Database \(TGAD\)](#).

**General Information** [Add to Basket](#) [View Basket](#)

Name: **Tb927.2.2900**  
 Systematic Name: Tb927.2.2900  
 Gene Synonyms: 10C8.250  
 Status: role inferred from homology  
 Products: ARP2/3 complex subunit, putative (5 others)  
 Type: CDS  
 Sequence: [DNA](#) and [Protein](#)

**Location**  
 Chromosome: 2  
 Chromosome Location: 554106..554648 Length: 543 bp  
[Graphical Display \(in Artemis\)](#) [Genome Browser](#)  
 Context Map:

**Primary Annotation**  
 ARP2/3 complex subunit 4 (curated by B. Wickstead, Univ. of Oxford)

**Predicted Peptide Properties**

Mass	20.7 kDa	Amino acids	180
Isoelectric point	pH 7.9	Charge	3.0

Signal Peptide: Not found  
 Transmembrane Domains: 0 found  
 GPI Anchor: Not found  
 Protein Map:

**Domain Information**

DB	Accs	Description	Note
Pfam	<a href="#">PF05856</a>	ARP2/3 complex 20 kDa subunit (ARPC4)	Residue: 1-177 (Score: 1.8e-60)
InterPro	<a href="#">IPR008384</a>	ARP2/3 complex 20 kDa subunit	Derived from hit: Pfam

**Gene Ontology Annotation**

Term (browse AmiGO)	Qualifier	Evidence	Other genes annotated to this term
<a href="#">Biological Process</a>			
<a href="#">actin filament polymerization</a>	ISS (TIGR_Tba1:Tb927.2.2900 TIGR_REF:GO_ref) with SWISS-PROT:O15509		<a href="#">1 other</a>
<a href="#">actin filament polymerization</a>	IEA (GO:interpro2go)		<a href="#">1 other</a>
<a href="#">Cellular Component</a>			
<a href="#">ARP2/3 protein complex</a>	ISS (TIGR_Tba1:Tb927.2.2900 TIGR_REF:GO_ref) with SWISS-PROT:O15509		<a href="#">2 others</a>
<a href="#">cytoskeleton</a>	IEA (GO:interpro2go)		<a href="#">126 others</a>
<a href="#">Molecular Function</a>			
<a href="#">structural constituent of cytoskeleton</a>	ISS (TIGR_Tba1:Tb927.2.2900 TIGR_REF:GO_ref) with SWISS-PROT:O15509		<a href="#">20 others</a>

**Literature**  
 Search for gene/protein in [PubMed](#)

**Database Cross-References**

DB	Accs	Description	Note
UniProt	<a href="#">Q7YVE2</a>	Actin-related protein, ARP2/3 complex subunit, putative.	
TGAD	<a href="#">Tb927.2.2900</a>	TIGR's T. brucei Annotation Database (TGAD)	

**Orthologues**

DB	Accs	Description	Note
GeneDB_Lmajor	<a href="#">LmjF02.0600</a>	ARP2/3 complex subunit, putative	predicted by jaccard cog clustering
GeneDB_Tcruzi	<a href="#">Tc00.1047053509127.104</a>	ARP2/3 complex subunit, putative	predicted by jaccard cog clustering
GeneDB_Tcruzi	<a href="#">Tc00.1047053508737.194</a>	ARP2/3 complex subunit, putative	predicted by jaccard cog clustering

**Database Similarities**

**AmiGO**

Search GO

Exact Match  
 Terms  
 Gene Symbol/Name  
 Submit Query  
 Advanced Query

all: all (3754)

- GO:0008150: biological\_process (3419)
- GO:0005575: cellular\_component (3074)
- GO:0003674: molecular\_function (3530)
- GO:0005198: structural\_molecule\_activity (193)
- GO:0005200: structural\_constituent\_of\_cytoskeleton (21)**
- obsolete\_biological\_process: obsolete\_biological\_process (1)
- obsolete\_cellular\_component: obsolete\_cellular\_component (1)
- obsolete\_molecular\_function: obsolete\_molecular\_function (1)

2 This local copy of the AmiGO browser provides access to proteins annotated to the same ontology term. Initially, you will see only the *T. brucei* proteins annotated to this term. Click on the hyperlinked term.

## 3

This is the top of the browser page showing the descriptions of the *T. brucei* proteins. Clicking on **1** will show you the other ontologies this particular protein is annotated to. **2** will take you to the feature page in GeneDB and the evidence code **TAS** (**3**) will take you to the paper substantiating this assignment.

**AmiGO** Last updated: 2005-08-28

**structural constituent of cytoskeleton**

**Accession:** GO:0005200  
**Aspect:** molecular\_function  
**Synonyms:** None  
**Definition:**  
 The action of a molecule that contributes to the structural integrity of a cytoskeletal structure.

**Term Lineage**

all : all ( 3754 )  
 GO:0003674 : molecular\_function ( 3530 )  
 GO:0005198 : structural molecule activity ( 193 )  
 GO:0005200 : **structural constituent of cytoskeleton ( 21 )**

**External References**

InterPro ( 3 )  
 Pfam ( 2 )  
 PRINTS ( 1 )  
 PROSITE ( 2 )  
 SP\_KW ( 1 )

**All Gene Product Associations** Get ALL associations here:

**5** **6** **4**

**Datasource** **Evidence Code** **Species**

FlyBase All S. avermitilis  
 SGD All Curator Approved S. cerevisiae  
 MGI IMP S. coelicolor

Gene Symbol	Type	Datasource	Evidence	Full Name
<b>1</b> structural constituent of cytoskeleton				
<input type="checkbox"/> 3850	gene	GeneDB_Tb	<b>2</b>	actin-like protein 3, putative
<input type="checkbox"/> 3960	gene	GeneDB_Tb	IEA	actin, putative
<input type="checkbox"/> 1.0620	gene	GeneDB_Tbrucei	IEA	actin A
<input type="checkbox"/> Tb09.211.0630	gene	GeneDB_Tbrucei	IEA	actin A
<input type="checkbox"/> Tb10.61.0500	gene	GeneDB_Tbrucei	IEA	actin-like protein 2, putative
<input type="checkbox"/> Tb10.70.5830	gene	GeneDB_Tbrucei	IEA	actin-like protein, putative
<input type="checkbox"/> Tb11.01.1870	gene	GeneDB_Tbrucei	IEA	actin-like protein, putative
<input type="checkbox"/> Tb927.1.2330	gene	GeneDB_Tbrucei	<b>3</b> TAS	alpha tubulin
<input type="checkbox"/> Tb927.1.2340	gene	GeneDB_Tbrucei	TAS	alpha tubulin
<input type="checkbox"/> Tb927.1.2350	gene	GeneDB_Tbrucei	TAS	alpha tubulin
<input type="checkbox"/> Tb927.1.2360	gene	GeneDB_Tbrucei	TAS	alpha tubulin

By changing the filter settings, you can retrieve proteins from other organisms annotated to this term as well either by species name (**4**) as or database (**5**) restricting the search to a term associated with a product via a particular evidence code (**6**). Try this by searching for cytoskeleton constituents in the yeast *S. cerevisiae*. The *S. cerevisiae* database is called SGD.

## 4

Select 'SGD' and/or *S. cerevisiae* from the pull down menu and press 'Submit Query'.

**All Gene Product Associations** Get ALL associations here:

**Datasource** **Evidence Code** **Species**

FlyBase All S. avermitilis  
 SGD All Curator Approved S. cerevisiae  
 MGI IMP S. coelicolor

Below are some of the proteins annotated as cytoskeletal components in *S. cerevisiae*. Click on the link to SGD and have a look at the status of annotation of this particular protein. Also, explore what happens if you were to filter on some of the other GO evidence codes such as **IDA**.

## AmiGO

Last updated: 2005-08-28

### structural constituent of cytoskeleton

**Accession:** GO:0005200  
**Aspect:** molecular\_function  
**Synonyms:** None  
**Definition:**  
 The action of a molecule that contributes to the structural integrity of a cytoskeletal structure.

#### Term Lineage

all : all ( 6457 )  
 GO:0003674 : molecular\_function ( 6451 )  
 GO:0005198 : structural molecule activity ( 356 )  
 GO:0005200 : **structural constituent of cytoskeleton ( 49 )**

#### External References

InterPro ( 3 )  
 Pfam ( 2 )  
 PRINTS ( 1 )  
 PROSITE ( 2 )  
 SP\_KW ( 1 )

#### All Gene Product Associations

Get ALL associations here:

All Associations With Terms

#### Filter Associations

Datasource	Evidence Code	Species
All	All	All
FlyBase	All Curator Approved	A. aeolicus
SGD	IMP	A. fulgidus

Gene Symbol	Type	Datasource	Evidence	Full Name
-------------	------	------------	----------	-----------

#### structural constituent of cytoskeleton

<input type="checkbox"/>	ACT1	gene	SGD	TAS	actin
<input type="checkbox"/>	ARC18	gene	SGD	TAS	None
<input type="checkbox"/>	ARC40	gene	SGD	TAS	None
<input type="checkbox"/>	ARP1	gene	SGD	IPI	None
<input type="checkbox"/>	ARP2	gene	SGD	TAS	actin related protein
<input type="checkbox"/>	ARP3	gene	SGD	TAS	None
<input type="checkbox"/>	ASK1	gene	SGD	IDA	None
<input type="checkbox"/>	BBP1	gene	SGD	IPI	None
<input type="checkbox"/>	BIM1	gene	SGD	IPI	None

### Exercise 1.5 The use of BLAST to identify sequences with similarity to known Arp2/3 complex subunits

In addition to using gene ontology assignments, you could have also used similarity searches to identify the two missing actin-related proteins. To identify the putative *T. brucei* Arp3 homologue, you could also use the BLAST tool integrated into GeneDB.

As mentioned previously, the components of the Arp2/3 complex are highly conserved amongst eukaryotes. Therefore, one approach to identifying putative homologues in *T. brucei* is to use the *S. pombe* Arp3 sequence as a query sequence in a BLAST search.

2 Retrieve the amino acid sequence by clicking on the sequence link.

### 3

Click on ‘**send to omniBLAST**’ link.  
omniBLAST permits searching across  
different databases selected by the user.

-100-



4

This is the omniBLAST page, providing access to databases of all sequences housed in GeneDB. By default, the datasets of the organism you started out from will be selected. For this search select the *S. pombe* and *T. brucei* protein databases. Note that the amino acid sequence of the *S. pombe* Arp3 protein has automatically been pasted into the query sequence box. Start omniBLAST by clicking on the 'Start omniBLAST' button.

GeneDB omniBLAST Server

Go To:  Search:  Help

Perform a BLAST search on a set of protein databases (BLASTP or BLASTX, depending on the query sequence) or nucleotide databases (BLASTN and TBLASTX or TBLASTN) and obtain a list of the best five HSP for each database. If there are any HSP you can click on Full Search to see the complete BLAST output.

Databases with different parameters:  single organism BLAST:

QUERY DATA

Enter text to do:

Determine sequence type automatically ☐ or set sequence type to DNA ☐ protein ☒

Note: omniBLAST searches may take several minutes depending on the number of selected databases. Please check the databases chosen below are correct.

AST databases selected below:

Jump down page to: [Fungi](#) [Bacteria](#) [Eukarya](#) [Eukarya/Archaea](#)

Applied BAC sequences ☒ *A. fumigatus* BAC predicted genes ☒ *A. fumigatus* BAC ends ☒

AC predicted genes ☒ *S. pombe* peptides ☒

Whole genome shotgun reads ☒ *T. brucei* predicted proteins ☒

Genome contigs ☒ *T. brucei* predicted genes ☒

504 sequences ☒ *T. brucei* predicted proteins ☒

Genome contigs 1.1-1.5 Mb region ☒ *T. brucei* predicted genes ☒

Genome contigs ☒ *T. brucei* predicted proteins ☒

GeneDB coding ☒ *T. brucei* predicted genes ☒

Major GeneDB (DNA) ☒ *T. brucei* predicted proteins ☒

*L. major* SPTX data ☒ *L. major* EMBL data ☒

All *Leishmania* species SPTX data ☒ All *Leishmania* species EMBL data ☒

*L. major* end sequences (Sanger Institute, SBB1 and WashU) ☒ *L. major* chromosome shotgun ☒

*L. major* unassembled shotgun reads ☒ *L. major* ESTs ☒

*L. infantum* contigs ☒ *L. infantum* unassembled shotgun reads ☒

*P. berghei* ☒ *P. berghei* GeneDB ☒

*P. chabaudi* ☒ *P. chabaudi* GeneDB ☒

*P. falciparum* ☒ *P. falciparum* selected chromosome sequences ☒

*P. falciparum* Sanger Reads ☒ *P. falciparum* proteins ☒

*T. annulata* ☒ *T. annulata* GeneDB ☒

*T. cruzi* ☒ *T. cruzi* GeneDB predicted genes (coding sequences) ☒

*T. cruzi* contigs ☒ *T. cruzi* predicted proteins ☒

*T. cruzi* EMBL data ☒ *T. cruzi* SPTX data ☒

*T. brucei* (More information on databases) ☒ *T. brucei* predicted genes (coding sequences) ☒

*T. brucei* contigs ☒ *T. brucei* predicted proteins ☒

*T. brucei* chromosome 10 reads ☒ *T. brucei* chromosome 9 reads ☒

African trypanosome EMBL data ☒ *T. brucei* chromosome 11 reads ☒

*T. brucei* GSEEST clusters ☒ African trypanosome SPTX data ☒

*T. brucei* ☒ *T. brucei* predicted genes (coding sequences) ☒

*T. brucei* contigs ☒ *T. brucei* predicted proteins ☒

*T. brucei* EMBL data ☒ *T. brucei* SPTX data ☒

*T. vivax* ☒ *T. vivax* GeneDB predicted genes (coding sequences) ☒

*T. vivax* reads ☒ *T. vivax* predicted proteins ☒

*T. vivax* ☒ *T. vivax* contigs ☒

*B. bronchiseptica* ☒ *B. bronchiseptica* complete sequence ☒

*B. bronchiseptica* ☒ *B. bronchiseptica* predicted proteins ☒

*B. paratuberculosis* ☒ *B. paratuberculosis* complete sequence ☒

*B. paratuberculosis* ☒ *B. paratuberculosis* predicted proteins ☒

*B. pertussis* ☒ *B. pertussis* complete sequence ☒

*B. pertussis* ☒ *B. pertussis* predicted proteins ☒

*S. typhi* ☒ *S. typhi* chromosomal sequence ☒

*S. typhi* ☒ *S. typhi* pHCM1 sequence ☒

*S. typhi* ☒ *S. typhi* pHCM2 sequence ☒

*S. typhi* ☒ *S. typhi* pHCM1 proteins ☒

*S. typhi* ☒ *S. typhi* pHCM2 proteins ☒

*G. moritatus* ☒ *G. moritatus* clustered ESTs ☒

RESULTS

5

Retrieve the results by clicking on the 'retrieve' button. As indicated, results will be accessible for the next 2 weeks using the listed URL.

GeneDB


The Wellcome Trust  
Sanger Institute  
Pathogen Sequencing UnitGo To:  Search:  HelpRetrieve result for id:  

Your BLAST query has been added to the queue of jobs.  
The majority of BLASTs are completed within two minutes.

To retrieve your results, click the **retrieve** button above, or use the following URL: <http://dev.gene-db.org/gene-db2/blast/getblast?id=s2bFXg8107h096508z39683>

## 6

You will retrieve an abbreviated BLAST results page, listing only the top 5 hits without alignments. Click on the '**Full BLAST Search**' of your query sequence against the predicted *T. brucei* proteome.

**GeneDB** **Blast Server Results** 

Go To:  Search:  [Help](#)

Retrieve result for id:

At peak times your BLAST searches could take longer than normal. Please be patient.

BLAST results are kept on our [servers](#) for three days following query submission. Results may be retrieved any number of times during this period. After this time queries must be resubmitted if further examination is required.

**Summary for: *S. pombe* proteins [wublastp], for query: SPAC630.03** [\[Full BLAST Search\]](#)

Name: <a href="#">SPAC630.03, arp3</a>	Score: 2229	(P/N): 6.3e-233	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">SPBC32H8.12c, act1</a>	Score: 505	(P/N): 9.3e-59	N: 2	<a href="#">[Full Sequence]</a>
Name: <a href="#">SPAC11H11.06, arp2</a>	Score: 507	(P/N): 1.2e-58	N: 2	<a href="#">[Full Sequence]</a>
Name: <a href="#">SPBC1347.12</a>	Score: 452	(P/N): 1.3e-44	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">SPBP23A10.08, alp5</a>	Score: 214	(P/N): 2.6e-29	N: 3	<a href="#">[Full Sequence]</a>

**Summary for: *T. brucei* predicted proteins [wublastp], for query: SPAC630.03** [\[Full BLAST Search\]](#)

Name: <a href="#">Tb09.160.3850</a>	Score: 970	(P/N): 3.0e-99	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">Tb09.211.0630</a>	Score: 493	(P/N): 1.0e-48	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">Tb09.211.0620</a>	Score: 493	(P/N): 1.0e-48	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">Tb10.61.0500</a>	Score: 420	(P/N): 5.7e-41	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">Tb11.01.1870</a>	Score: 308	(P/N): 5.1e-34	N: 2	<a href="#">[Full Sequence]</a>

8

BLAST results are kept on our servers for three days following query submission. Results may be retrieved any number of times during this period. After this time queries must be resubmitted if further examination is required.

Low complexity filtering disabled  
Repeatmasker disabled

BLAST 2.0M2-WashU [15-Sep-2002] [default.x64-03-EZLFP64 2002-09-18T19:28:12]

Copyright (C) 1994-2002 Washington University, Saint Louis, Missouri USA.  
All Rights Reserved.

Reference: Gish, W. (1996-2002) <http://hlust.wustl.edu>

Query: SFAC630.03  
(477 letters)

Database: TXFP1910.gnats  
16,376 sequences; 5,322,143 total letters.  
Searching...10...20...30...40...50...60...70...80...90...100% done

9

**GeneDB** CDS: Tb09.160.3850 **GeneDB**

**Please Note:** Over the Christmas period, some feedback and GNC submissions were lost due to technical problems. We have used our log files to trace these and created a list of fixes, with appropriate dates. If feedback you have sent us in the list, could you please re-submit it, should it still be appropriate.

Search for:  Go To: Organisms  Go To: Shortcuts  [Help](#) [Contact us](#)

---

**General Information** [Add to Favourites](#) [View Results](#)

<b>Name</b>	Tb09.160.3850 <span style="color: red;">(Warning: Temporary systematic id)</span>
<b>Gene Synonym</b>	20216.235
<b>Protein Synonym</b>	chr20c0g160.tsp0290
<b>Notes</b>	role inferred from homology
<b>Products</b>	actin-related protein 3, putative
<b>Type</b>	CDS
<b>Sequence</b>	<a href="#">DNA</a> and <a href="#">Protein</a>

---

**Location**

<b>Chromosome</b>	9
<b>Contig Location</b>	853771..855021 Length: 1251 bp

[Graphical Display \(in Artemis\)](#)

**Context Map:**

[Tb09.160.3773](#) [Tb09.160.3780](#) [Tb09.160.3790](#) [Tb09.160.3800](#) [Tb09.160.3810](#) [Tb09.160.3820](#) [Tb09.160.3830](#) [Tb09.160.3840](#) **[Tb09.160.3850](#)** [Tb09.160.3859](#) [Tb09.160.3870](#)  
[Tb09.160.3880](#) [Tb09.160.3890](#) [Tb09.160.3900](#) [Tb09.160.3910](#) [Tb09.160.3920](#) [Tb09.160.3930](#) [Tb09.160.3940](#) [Tb09.160.3950](#) [Tb09.160.3960](#) [Tb09.160.3970](#) [Tb09.160.3980](#)

---

**Primary Annotation**

Wbly role in F-actin nucleation as part of complex with ABP2 (inferred by B. Wickstead, Univ. of Oxford)

---

**Predicted Peptide Properties**

<b>Mass</b>	47.7 kDa	<b>Amino acids</b>	436
<b>Isoelectric point</b>	pI 6.4	<b>Charge</b>	-1.0

---

**Signal Peptide** Not found

**Transmembrane Domains** 0 found

**GPI Anchor** Not found

**Protein Map:**

**Domain Information:**

DB	Access	Description	Note
Pfam	<a href="#">PF00022</a>	Actin	Residue 4-410 (Score 36-38)
InterPro	<a href="#">IPR010400</a>	Actin/actin-like	Derived from InterPro SMART PF00022 Pfam
PROSITE	<a href="#">PS00120</a>	Actin signature	Residue 49-40 (Score 2.4e-11), 84-102 (2.4e-11) and 152-171 (2.4e-11)
SMART	<a href="#">SM00104</a>	SMART	Residue 3-410 (Score 3 to 140)

---

**Gene Ontology Annotation**

Term (Access ID)	Qualifier	Evidence	Other genes associated in this term
<b>Cellular Component</b>			
<a href="#">actin cytoskeleton</a>		ISA from InterPro/GO mapping	<a href="#">13 others</a>
<b>Molecular Function</b>			
<a href="#">structural constituent of cytoskeleton</a>		ISA from InterPro/GO mapping	<a href="#">13 others</a>

---

**Literature**

---

**Orthologues**

DB	Access	Description	Note
GenBank	<a href="#">U00167.1</a>	actin-related protein 3, putative	predicted by jvarkit gene clustering
GenBank	<a href="#">U00167.1</a>	actin-related protein 3, putative	predicted by jvarkit gene clustering

---

**Paralogous Family**

---

**Database Similarities**

DB	Access	Organism	Description	Value	Overlap	Algorithm
TrEMBL	<a href="#">A02212</a>	<i>Chlamydomonas reinhardtii</i>	actin-like protein 3	2.5e-89	91.94%	415 aa
TrEMBL	<a href="#">A02212</a>	<i>Chlamydomonas reinhardtii</i>	actin-like protein 3	2.0e-89	92.73%	421 aa

10

[illegible]

	<i>S. pombe</i>	<i>T. brucei</i>	
<b>Arp2</b>	SPAC11H11.06	Tb10.61.0500	
<b>Arp3</b>	SPAC630.03		
<b>p41-Arc</b>	SPBC14C8.06		
<b>p34-Arc (PF04045)</b>	SPAC6F6.10c		
<b>p21-Arc (PF04062)</b>	SPBC1778.08c	Tb10.70.2680	
<b>p20-Arc (PF05856)</b>	SPAC6G9.07c	Tb927.2.2900	
<b>p16-Arc (PF04699)</b>	SPAC17G8.04c		

Using 4 different approaches to retrieve/identify putative homologues, you should have completed this table. As you will have noticed, you probably wouldn't have been able to retrieve all the data by just using a single approach to mine the *T. brucei* genome, which highlights some of the issues outlined in the introduction to this module.

### Exercise 1.6 Identify the Arp2/3 complex in other kinetoplastid species

Imagine now that you are not only interested in this complex in *T. brucei* but also in other *Trypanosoma* and *Leishmania*, causative agents of Leishmaniasis, species. GeneDB is ideally suited to this purpose as it houses sequence and annotation of multiple kinetoplastid species and the data are extensively cross-linked. You could start by identifying components in *L. major* and *T. cruzi* and then move on to *L. infantum* and the cattle-infective *T. vivax* and *T. congolense*.

There are a number of ways you could tackle this problem. You could use similarity searches, GO and/or Pfam catalogues, similar to what you have been doing in the previous exercises. However, a faster way would be to make use of orthologue cross-links provided by GeneDB.

1

There are two ways you can do this, either by using the orthologue cross-links provided on each of the gene pages or alternatively, use the 'List Download' option. Start with having a look at the orthologue cross-links on the gene page. Go to the gene page showing annotation associated with Tb927.2.2900.

**GeneDB** CDS: Tb927.2.2900

WARNING: 5th September 2005. The Full Context Search is still pointing at the previous data run's data. This will be fixed later in the day.

Search for:  Go To:  Organisms  Shortcuts  Help  Contact creator

The sequence and annotation on this page were provided by [TIGR's T. brucei project](#).  
This gene is in [TIGR's T. brucei Annotation Database \(TGAD\)](#)

**General Information** [Add to Basket](#) [View Basket](#)

Name: Tb927.2.2900  
Systematic Name: Tb927.2.2900  
Gene Synonym: 10C8.250  
Status: role inferred from homology  
Products: ABP23 complex subunit, putative (2 others)  
Type: CDS  
Sequence: [DNA](#) and [Protein](#)

**Location**

Chromosome: 2  
Chromosome Location: 554106..554648 Length: 543 bp  
[Graphical Display \(in Artemis\)](#) [Genome Explorer](#)

Context Map:

[Tb927.2.2830](#) [Tb927.2.2860](#) [Tb927.2.2880](#) **Tb927.2.2900** [Tb927.2.2910](#) [Tb927.2.2920](#) [Tb927.2.2940](#) [Tb927.2.2950](#) [Tb927.2.2970](#)

**Primary Annotation**

ABP23 complex subunit 4 (curated by B. Wickstead, Univ. of Oxford)

**Predicted Peptide Properties**

Mass	20.7 kDa	Amino acids	180
Isoelectric point	pH 7.9	Charge	3.0

Signal Peptide: Not found  
Transmembrane Domains: 0 found  
GPI Anchor: Not found  
Protein Map:

**Domain Information**

ID	Accession	Description	Note
B	PF00306	ABP23 complex 20 kDa subunit (ABP23)	Residue: 1-177 (Score: 1.8e-60)
InterPro	IPR003384	ABP23 complex 20 kDa subunit	Derived from InterPro

**Gene Ontology Annotation**

Term (Accession)	Qualifier	Evidence	Other genes annotated to this term
actin filament polymerization (GO:0000719)	ISS	(TIGR, Tb927.2.2900(TIGR_REF_GO_ref) with SWISS-PROT O15509)	1 other
actin filament polymerization (GO:0000719)	IEA	(GO:0000719)	1 other
ABP23 complex (GO:0000719)	ISS	(TIGR, Tb927.2.2900(TIGR_REF_GO_ref) with SWISS-PROT O15509)	2 others
ABP23 complex (GO:0000719)	IEA	(GO:0000719)	126 others
cytoskeleton (GO:0000719)	ISS	(TIGR, Tb927.2.2900(TIGR_REF_GO_ref) with SWISS-PROT O15509)	20 others

**Literature**

Search for gene/protein in [PubMed](#)

**Database Cross-References**

ID	Accession	Description
UniProt	Q7TVE2	Actin-related protein, ABP23 complex subunit, putative
TGAD	Tb927.2.2900	TIGR's T. brucei Annotation Database (TGAD)

**Orthologues**

ID	Accession	Description	Note
GeneDB	LmjF02.0600	ABP23 complex subunit, putative	predicted by jaccard cog clustering
GeneDB	Tm00.1047035509127.104	ABP23 complex subunit, putative	predicted by jaccard cog clustering
GeneDB	Tm00.1047035509127.104	ABP23 complex subunit, putative	predicted by jaccard cog clustering

**Database Similarities**

2

This part of the gene page provides links to manually curated orthologues in other species.

3

Click on the link providing access to the putative *L. major* orthologue.

**GeneDB** CDS: LmjF02.0600

Please Note: Over the Christmas period, some feedback and GNC submissions were lost due to technical problems. We have used our log files to trace these and curated a list of them, with appropriate dates. If feedback you have sent is on this list, could you please re-submit it, should it still be appropriate.

Search for:  Go To:  Organisms  Shortcuts  Help  Contact creator

These sequence data were generated by [SRA](#)

**General Information** [Add to Basket](#) [View Basket](#)

Name: LmjF02.0600  
Systematic Name: LmjF02.0600  
Prev. Systematic Id: LmjF02.0600  
Status: function inferred from homology  
Products: ABP23 complex subunit, putative (2 others)  
Type: CDS  
Sequence: [DNA](#) and [Protein](#)

**Location**

Chromosome: 2 ([Chromosome 2 Sequencing Project](#))  
Contig: LmjF02\_01\_20040630\_V4.0  
Location: 313087..313803 Length: 717 bp  
[Graphical Display \(in Artemis\)](#)

Context Map:

[LmjF02.0580](#) [LmjF02.0590](#) **LmjF02.0600** [LmjF02.0610](#) [LmjF02.0620](#) [LmjF02.0630](#) [LmjF02.0640](#)

**Primary Annotation**

predicted by codon-gram to code glaser hexamer  
ORF = (312964..313803)



4

You could now go through each one of the 7 putative members of the *T. brucei* Arp2/3 complex, identifying putative orthologues by looking at the ‘**Orthologues**’ section on the gene page. There is however a faster way using the ‘**List Download**’ utility. This function allows you to compile a list of your gene of interest and then subsequently downloading the description and sequence of these features using the ‘**Gene Basket**’.

Start by going to the top of the gene page. Imagine this to be your first gene of interest. In order to collect your genes of interest, you’ll need to click on the ‘**Add to Basket**’ icon at the top of the page. This will now have added the identifier of this gene to the virtual basket.

The screenshot shows the GeneDB website for the gene CDS: Tb927.2.2900. The page includes a search bar, a navigation bar with 'Add to Basket' and 'New Basket' buttons, and a table of gene properties.

General Information	
Name	Tb927.2.2900
Systematic Name	Tb927.2.2900
Gene Synonyms	3008.250
Protein	role inferred from homology
Products	Arp2/3 complex subunit, cytosolic (UniProt)
Type	CDS
Sequence	FASTA and Protein

Location	
Chromosome	2
Chromosome Location	554016-554648 Length 642 bp

Context Map:

Primary Annotation	
Arp2/3 complex subunit 4 (annotated by B. Wickham, Univ. of Oxford)	

Predicted Protein Properties	
Mass	20.7 kDa
Isoelectric point	pI 5.73
Amino acids	180
Charge	3.0

5

Now go to each one of the gene pages of the putative Arp2/3 complex members – they are:

Tb10.70.2680  
 Tb10.61.0500  
 Tb10.389.0270  
 Tb10.406.0320  
 Tb927.2.2900  
 Tb927.8.4410  
 Tb09.160.3850

and should all be listed in the table you filled in earlier. You can simply navigate between gene pages by filling in the ‘**Search for**’ box in the navigation bar. Add them to the gene basket by simply clicking on the ‘**Add to Basket**’ icon at the top of each of the gene pages.

6

Once you have added all your gene of interest to the basket, click on the ‘View Basket’ icon.

**GeneDB** T. brucei **GeneDB**

**CDS: Tb09.160.3850**

**Please Note:** Over the Christmas period, some feedback and GNC submissions were lost due to technical problems. We have used our log files to trace these and curated a [list](#) of them, with appropriate dates. If feedback you have sent is on this list, could you please resubmit it, should it still be appropriate.

Search for:  Go To: Organisms  Go To: Shortcuts  [Help](#) [Contact curator](#)

**General Information** [Add to Basket](#) [View Basket](#)

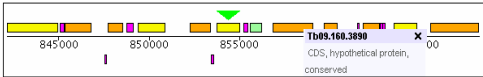
**Name**  
**Systematic Name** Tb09.160.3850 (Warning: Temporary systematic id)  
**Gene Synonyms** 28G16.235  
**Prev. Systematic Id** chr9:contig160.tmp0298  
**Status** role inferred from homology  
**Products** actin-related protein 3, putative  
**Type** CDS  
**Sequence** [DNA](#) and [Protein](#)

**Location**

**Chromosome** 9  
**Contig Location** 853771..855021 Length: 1251 bp

**Context Map:**

[Graphical Display \(in Artemis\)](#)



[MSH3](#) [Tb09.160.3770](#) [Tb09.160.3780](#) [Tb09.160.3790](#) [Tb09.160.3800](#) [Tb09.160.3810](#) [Tb09.160.3820](#) [Tb09.160.3830](#) [Tb09.160.3840](#) [Tb09.160.3850](#) [Tb09.160.3860](#) [Tb09.160.3870](#) [Tb09.160.3880](#) [Tb09.160.3890](#) [Tb09.160.3900](#) [Tb09.160.3910](#) [Tb09.160.3920](#) [Tb09.160.3930](#) [Tb09.160.3940](#) [Tb09.160.3950](#) [Tb09.160.3960](#) [Tb09.160.3970](#) [Tb09.160.3980](#)

**Primary Annotation**

likely role in F-actin nucleation as part of complex with ARP2 (curated by B. Wickstead, Univ. of Oxford)

**Predicted Peptide Properties**

<b>Mass</b>	47.3 kDa	<b>Amino acids</b>	416
<b>Isoelectric point</b>	pH 6.4	<b>Charge</b>	-1.0
<b>Signal Peptide</b>	Not found		

7

The page lists all your genes of interest in the gene basket.

**GeneDB** The Wellcome Trust  
Sanger Institute  
Pathogen Sequencing Unit

**Gene Basket**

Go To: Organisms  Go To: Shortcuts  [Help](#)

All of 7 results shown

[Report Download](#)

<i>T. brucei</i>	CDS	<a href="#">Tb10.70.2680</a>	ARP2/3 complex subunit, putative, ARP2/3 complex subunit, putative, TRYP_x-70a06.p2kb545_360
<i>T. brucei</i>	CDS	<a href="#">Tb927.2.2900</a>	ARP2/3 complex subunit, putative, 10C8.250
<i>T. brucei</i>	CDS	<a href="#">Tb927.8.4410</a>	ARP2/3 complex subunit, putative, Tb08.29H22.800
<i>T. brucei</i>	CDS	<a href="#">Tb09.160.3850</a>	actin-like protein 3, putative, 28G16.235, chr9:contig160.tmp0298, obsolete:Chr9 tmp.121c, obsolete:chr9_contig3519 tmp.166c
<i>T. brucei</i>	CDS	<a href="#">Tb10.389.0270</a>	actin related protein 2/3 complex, putative, ARP2/3 complex subunit, putative
<i>T. brucei</i>	CDS	<a href="#">Tb10.61.0500</a>	actin-like protein 2, putative, TRYP_x-61h03.q2kb185_78
<i>T. brucei</i>	CDS	<a href="#">Tb10.406.0320</a>	ARP2/3 complex 16kDa subunit, putative, ARP2/3 complex subunit, putative

Hosted by the [Sanger Institute](#) [Send comments, requests, corrections and updates](#)

8

Click on the ‘Report Download’ link.

9

This page allows you to download the description, sequence (both nucleotide and amino acid) and more importantly, the orthologues of your gene of interest.

Go To **Organisms** Go To **Shortcuts** [Help](#)

**ID List**

Please enter your gene/ma ids either:

- as database cross-references (eg in the form GeneDB\_Spombe:SPAC1002.09c).
- or you can just use the systematic ids eg (Tb927.1.710) **but** in this case you must also set the default organism.

Default Organism: **T. brucei**

Tb10.61.0500  
Tb10.70.2680  
Tb10.389.0270  
Tb10.406.0320  
Tb08.29H22.800  
Tb09.160.3850  
Tb927.2.2900

**Information Required**

☐ Descriptions  
☐ DNA (Unspliced sequence) or cDNA  
☐ DNA (Spliced sequence)  
☐ Protein sequence  
☐ Intergenic Sequence (3')  
☐ Intergenic Sequence (5')  
☐ Sequence Range  
☒ Orthologues

Number of bases: **20**  
 3' distance: **0** 5' distance: **0**

11

The browser window return the annotated orthologues in *L. major* and *T. cruzi*.

T. brucei	Tb09.160.3850	GeneDB_Tcruzi:Tc00.1047053508277.260, GeneDB_Tcruzi:Tc00.1047053503913.20
T. brucei	Tb927.8.4410	GeneDB_Lmajor:LmjF10.1000, GeneDB_Tcruzi:Tc00.1047053506865.10, GeneDB_Tcruzi:Tc00.104705350
T. brucei	Tb10.70.2680	GeneDB_Tcruzi:Tc00.1047053508625.30, GeneDB_Tcruzi:Tc00.1047053510963.70
T. brucei	Tb927.2.2900	GeneDB_Lmajor:LmjF02.0600, GeneDB_Tcruzi:Tc00.1047053509127.104, GeneDB_Tcruzi:Tc00.10470535
T. brucei	Tb10.406.0320	GeneDB_Tcruzi:Tc00.1047053511635.30
T. brucei	Tb10.389.0270	GeneDB_Lmajor:LmjF18.0920, GeneDB_Tcruzi:Tc00.1047053504215.40, GeneDB_Tcruzi:Tc00.104705351
T. brucei	Tb10.61.0500	GeneDB_Tcruzi:Tc00.1047053508899.110, GeneDB_Tcruzi:Tc00.1047053511361.40

12

Have a look at the results.

- Is *L. major* predicted to contain all components of the Arp2/3 complex?
- How many copies of each of the components is *T. cruzi* genome predicted to encode?
- Why might you find them in duplicates (hint: have a look at the *T. cruzi* GeneDB homepage)?
- How would you go about identifying homologues in the other *Leishmania* and *Trypanosoma* species?
- Send a feedback e-mail to the curators in cases where you come across mis-annotation.

	<i>T. brucei</i>	<i>L. major</i>	<i>T. cruzi</i>	<i>L. infantum</i>	<i>T. vivax</i>	<i>T. coongolense</i>	<i>T. b. gambiense</i>
<b>Arp2</b>	Tb10.61.0 500						
<b>Arp3</b>	Tb09. 160.3850						
<b>p41-Arc</b>	Tb10.380. 0270						
<b>p34-Arc (PF04045)</b>	Tb927.8.4 410						
<b>p21-Arc (PF04062)</b>	Tb10.70.2 680						
<b>p20-Arc (PF05856)</b>	Tb927.2. 2900						
<b>p16-Arc (PF04699)</b>	Tb10.406. 0320						

During this exercise you will have become familiar with GeneDB, the way data are displayed on feature pages and the various ways data can be accessed. As you will have seen, you wouldn't have been able to retrieve all the data by just using a single approach to mine the genomes, but that instead you needed to employ multiple search strategies. You will also have seen how to compile lists of genes of interest and how to download them for further examination/experimentation. Lastly, with the increasing emphasis on comparative genomics, you hopefully saw how GeneDB allows you to easily retrieve genes from related organisms.

## Exercise 2 Use of the Artemis Applet

As you will have seen earlier, GeneDB has an integrated Artemis applet. You'll be using the applet to view additionally annotated sequence features as well as to download a range of sequence and features.

The *T. brucei* genome contains 3 copies of the gene encoding phosphoglycerate kinase. The 3 isozymes are differentially expressed during the life cycle of *T. brucei* and are targeted to different organelles. Does the sequence provide you with any clues why this may be?

1

Find out where the 3 genes are located by typing '**phosphoglycerate kinase**' into the search box on the *T. brucei* homepage.

2

The results page indicates that the 3 genes are tandemly arrayed on chromosome I. Click on the hyperlinked Tb927.1.700 gene name to get to the feature page.

3

Click on '**Graphical Display (in Artemis)**' to open up the range download page

The screenshot shows the GeneDB interface for the gene Tb927.1.700. The 'Basic Information' section lists the gene name, systematic name, gene synonyms, state, product, type, and sequence. The 'Location' section shows the chromosome (I), chromosome location, and a graphical map. The 'Sequence' section shows the sequence and a graphical map. The 'Annotations' section lists various annotations, including protein domains, protein families, and protein structures. A red box highlights the 'Graphical Display (in Artemis)' link under the 'Location' section. An arrow points from this link to the 'Graphical Display (in Artemis)' link in the 'Location' section.



The range download page allows you to define the range of sequence you'd like to download in either EMBL or FASTA format or alternatively, open up in an Artemis applet. By default, the sequence 10kb upstream and downstream of your feature of interest will be selected.

GeneDB Range Download Page

General Information

Systematic id: Tb927.1.700  
 Location: complement(232504..233826)  
 Organism: *T. brucei*  
 Type: CDS  
 Contig: Tb927.1 (Length: 1064672 bp)

Range Options

Up/Down Stream  
 Upstream: 10000 Downstream: 10000

From/To  
 From: 22504 To: 23826

Download Types

☐ Download (features and sequence in EMBL format)  
☐ Download (sequence in FASTA format)  
☒ Artemis Applet ([Help on Artemis](#))

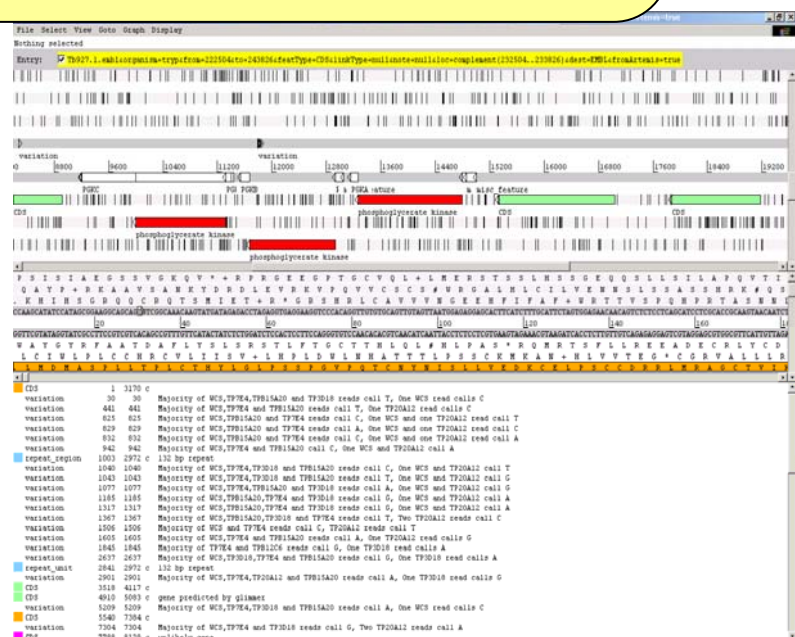
Please note the Artemis applet doesn't work on all browser/OS combinations. If it doesn't run for you please [install Artemis locally](#).

4

Select the Artemis applet and click on the 'Submit Query' button.

5

Find your region of interest in the applet. You'll see that in addition to the coding sequences having been annotated, the 5' and 3' UTRs have also been annotated. The 3' UTRs in particular, have been implicated in the differential regulation of mRNA stability at different life cycle stages in Kinetoplastids. In order to see whether this may also be the case here, you'll be downloading the sequence from the applet and aligning them using an alignment programme called clustalx.



6

Go back to the range download page and select the option to download the sequence in EMBL format. Press the 'Submit Query' button.

**GeneDB Range Download Page**

**General Information**

Systematic id: Tb927.1.700  
 Location: complement(232504..233826)  
 Organism: *T. brucei*  
 Type: CDS  
 Contig: Tb927.1 (Length: 1064672 bp)

**Range Options**

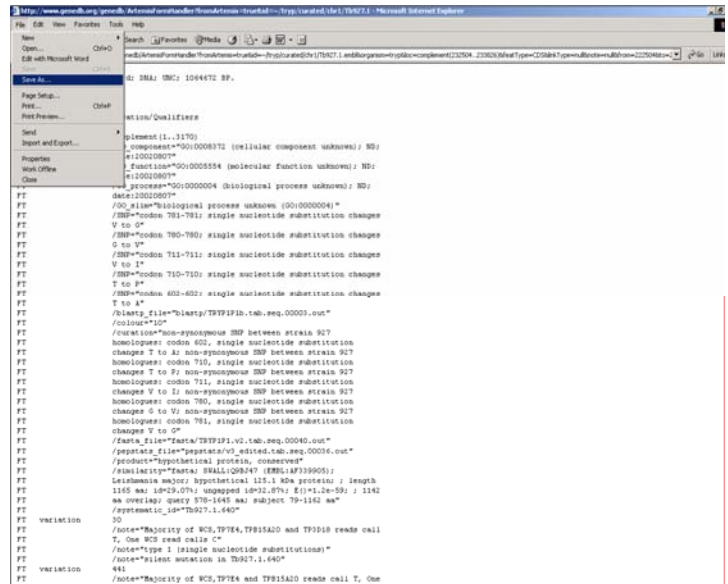
Up/Down Stream  
 Upstream: 10000 Downstream: 10000  
 From/To  
 From: 222504 To: 243826

**Download Types**

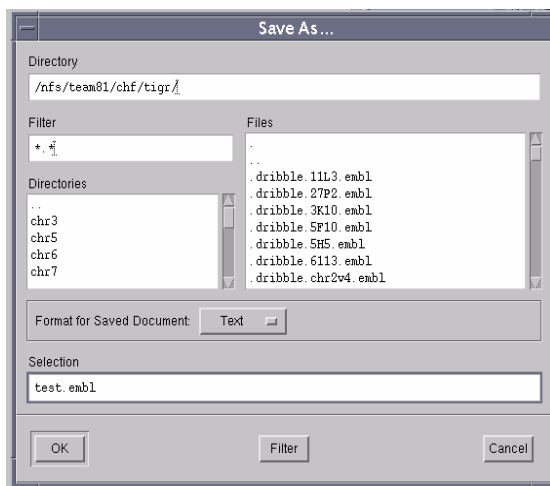
☒ Download (features and sequence in EMBL format)  
☐ Download (sequence in Fasta format)  
☐ Artemis Applet ([Help on Artemis](#))  
 Please note the Java applet doesn't work on all browsers/OS combinations. If it doesn't run for you please [install Artemis locally](#)

7

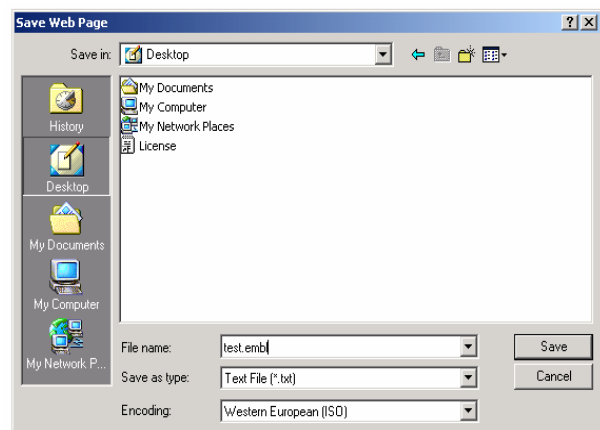
Save the sequence as a text file using your browser menu.



Unix/Linux



Windows



The screenshot shows the 'Open' menu of the 'Untitled' application window, which is highlighted with a red border. The menu options are 'File', 'Options', 'Open ...', 'Open from EBI ...', and 'Quit'. Below the menu, the text 'Genome Research' is visible. A file selection dialog titled 'Select a file...' is open, showing a list of files in the directory '/nfs/team81/chf/tigr/'. The file 'Tb927test.embl' is selected. The dialog also shows a list of folders on the left and a text field for the file name at the bottom.

9

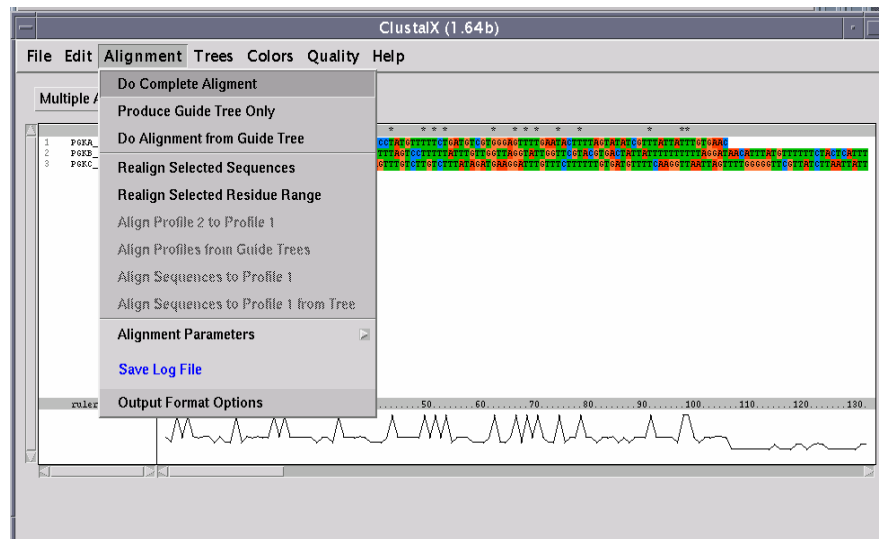
[illegible]

10

[illegible]

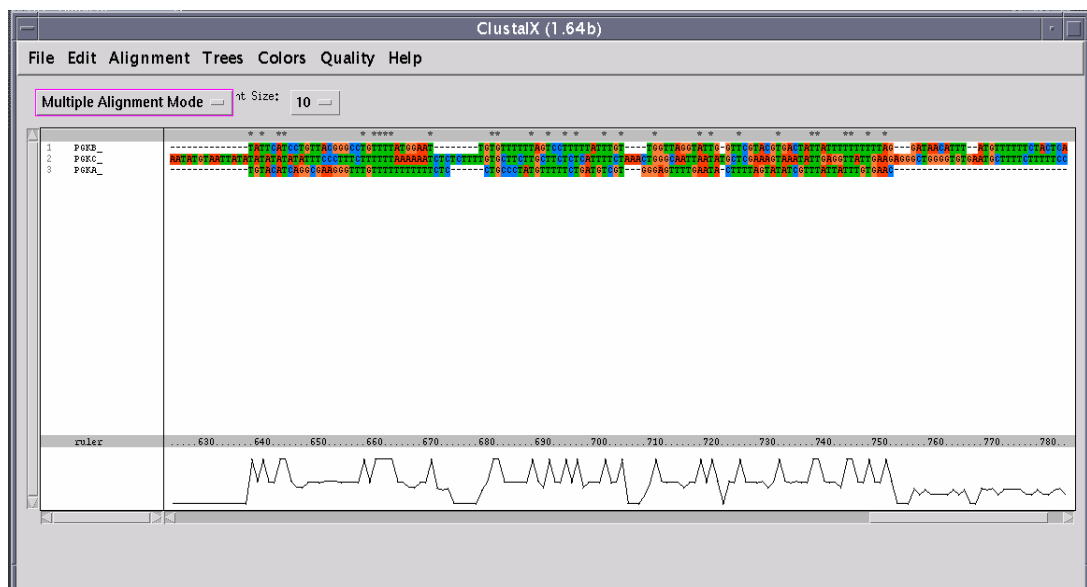
11

Change the alignment to 'Do Complete Alignment'.



12

Look at the alignment. Do the differences in UTR length and sequence tie in with the information you can get from the GeneDB pages regarding the localisation and expression of these 3 isozymes? Have also a look at the protein sequence alignments.



If you haven't got access to Artemis installed in the Unix/Linux environment, then you could always run the alignments using clustal via the web:  
<http://www.ebi.ac.uk/clustalw/>

### Exercise 3 Boolean querying

By now, you will have familiarised yourself with a variety of tools to search and browse the data housed in GeneDB. An additional query interface supports a wide range of queries on sequences and (curated) annotations stored in the relational database GUS. Searches can be combined with the boolean operators AND and OR. For example, users can select all proteins of a specified length range with a specified number of introns. Other query options include GO assignments, keywords, chromosome, protein domains and predicted protein sequence features. The queries in each session are tracked via a history page, allowing further refinement of searches and downloading of results as a nucleotide or amino acid FASTA file. This exercise will demonstrate how to combine/build up queries to retrieve a subset of predicted ABC transporters containing 8 transmembrane domains.

1

Select the link to the boolean querying interface by clicking on the 'complex querying pages' button on the *T. brucei* homepage.

2

Start off with querying the Pfam domain distribution.





3

Select the Pfam domain '**ABC transporter**' to retrieve proteins with this predicted domain

GeneDB

Go To: GeneDB Search: Simple Help

This page allows you to build more complex queries against the database using a preset selection of query forms and the boolean operators "AND" and "OR". For example, to construct a simple query begin by clicking on the "AND" button. Then use the pulldown menus to select two different queries. Finally, click on "proceed to next step" to generate a page that will allow you to specify parameters for the two queries. Running the resulting boolean query will return only those objects that satisfy query 1 AND query 2. **Please note:** you must select a query from each pulldown menu before the "proceed" button will work. If you need to back up a step, use the browser's "back" button.

You are currently searching *T. brucei* (Choose a different organism)

Proteins containing a specific Pfam domain

Pfam domain: **ABC transporter**

Query options

Rows per page: 20

Run query

Submit -- Reset

4

This is the results list.

Query: (Genes for *tryp*) intersect (Proteins that contain Pfam domain 'ABC transporter')

name	id	organism	description	
1	<a href="#">Tb027.1.4420</a>	<i>T. brucei</i>	ABC transporter, putative	Manual Annotation
2	<a href="#">Tb027.2.5410</a>	<i>T. brucei</i>	ABC transporter family protein (multidrug resistance), putative	Manual Annotation
3	<a href="#">Tb027.3.6130</a>	<i>T. brucei</i>	ABC transporter family protein (multidrug resistance), putative	Manual Annotation
4	<a href="#">Tb04.1H19.1030</a>	<i>T. brucei</i>	multidrug resistance-associated protein, putative	Manual Annotation
5	<a href="#">Tb04.1D20.510</a>	<i>T. brucei</i>	hypothetical protein, conserved	Manual Annotation
6	<a href="#">Tb09.160.4600</a>	<i>T. brucei</i>	multidrug resistance protein E	Manual Annotation
7	<a href="#">Tb09.28C22.830</a>	<i>T. brucei</i>	ABC transporter, putative	Manual Annotation
8	<a href="#">Tb10.61.0840</a>	<i>T. brucei</i>	ATP transporter, putative	Manual Annotation
9	<a href="#">Tb10.6K15.2900</a>	<i>T. brucei</i>	ABC transporter, putative	Manual Annotation
10	<a href="#">Tb10.6K15.3320</a>	<i>T. brucei</i>	ATP-binding cassette protein, putative	Manual Annotation
11	<a href="#">Tb10.70.4330</a>	<i>T. brucei</i>	ABC transporter protein, putative	Manual Annotation
12	<a href="#">Tb10.70.6290</a>	<i>T. brucei</i>	ribonuclease L inhibitor, putative	Manual Annotation
13	<a href="#">Tb06.5F5.130</a>	<i>T. brucei</i>	ABC transporter, putative	Manual Annotation
14	<a href="#">Tb08.26N11.730</a>	<i>T. brucei</i>	multidrug resistance protein A	Manual Annotation
15	<a href="#">Tb08.26A17.190</a>	<i>T. brucei</i>	hypothetical protein, conserved	Manual Annotation
16	<a href="#">Tb11.212.0110</a>	<i>T. brucei</i>	ABC transporter, putative	Manual Annotation
17	<a href="#">TRYP_m1062809.p1k.11</a>	<i>T. brucei</i>	Mitochondrial half-ABC transporter, predicted by automatic BLAST annotation	Automatic Annotation
18	<a href="#">TRYP_m1104609.p1k.38</a>	<i>T. brucei</i>	Tec114-2.11, predicted by automatic BLAST annotation	Automatic Annotation
19	<a href="#">TRYP_m162a11.p1k.2</a>	<i>T. brucei</i>	Alec10 protein, predicted by automatic BLAST annotation	Automatic Annotation
20				

To view this result set later, or to combine with others visit the [history page](#).

[Start a completely new complex query](#)

5

To identify only the predicted ABC transporters with 8 transmembrane domains, select the '**Start a completely new complex query**' from the bottom of the page.

6

Select '**T. brucei**' as an organism. Note, that you can also use the boolean querying interface to search across organism datasets.

GeneDB

GeneDB Boolean Search Page

Go To: GeneDB Search: Simple Help

Please choose which organism(s) you wish to search:

**Fungi**

☐ *A. fumigatus* ☐ *S. cerevisiae* ☐ *S. pombe*

**Protozoa**

☐ *D. discoideum* ☐ *E. histolytica* ☐ *E. tenella* ☐ *L. infantum*

☐ *E. major* ☐ *P. berghei* ☐ *P. chabaudi* ☐ *P. falciparum*

☐ *E. knowlesi* ☐ *T. annulata* ☐ *T. brucei* ☐ *T. congolense*

☐ *T. cruzi* ☐ *T. gambiense* ☐ *T. vivax*

**Parasite Helminths**

☐ *S. mansoni*

**Bacteria**

☐ *E. bronchiseptica* ☐ *E. fragilis* ☐ *E. paratuberculosis* ☐ *E. pertussis*

☐ *E. pseudomallae* ☐ *C. abortus* ☐ *C. diphtheriae* ☐ *E. carotovora*

☐ *S. aureus* MRSA ☐ *S. aureus* MSSA ☐ *S. coelicolor* ☐ *S. typhi*

**Parasite Vectors**

☐ *G. moritans*

**Viruses**

☐ *E. huxleyi virus* 86

7

Select 'Protein containing one or more predicted transmembrane domains' from the pull down menu.

The screenshot shows the GeneDB search interface. At the top, there is a search bar with 'Go To' set to 'GeneDB' and a 'Search' button. Below this, a message states: 'This page allows you to build more complex queries against the database using a preset selection of query forms and the boolean operators "AND" and "OR". For example, to construct a simple query begin by clicking on the "AND" button. Then use the pull-down menus to select two different queries. Finally, click on "proceed to next step" to generate a page that will allow you to specify parameters for the two queries. Running the resulting boolean query will return only those objects that satisfy query 1 AND query 2. Please note: you must select a query from each pull-down menu before the "proceed" button will work. If you need to back up a step, use the browser's "back" button.'

Below the message, it says 'You are currently searching *T. brucei*' with a link '(Choose a different organism)'. A section titled 'Choose a boolean condition or select a query' contains a dropdown menu. The menu is open, showing options like 'None', 'Proteins with a product containing a particular keyword or phrase', 'Proteins with an annotation matching a particular keyphrase', 'Proteins with a predicted GO function', 'Proteins with a predicted GO process', 'Proteins with a predicted GO component', 'Proteins within a range of length in amino acids', 'Proteins within a particular range of molecular masses', 'Proteins containing a specific Pfam domain', 'Proteins with a Pfam domain', 'Proteins containing a predicted signal peptide', and 'Protein containing one or more predicted transmembrane domains'. An arrow points from the instruction box to this option.

At the bottom right, there is a link 'Send us your comments on GeneDB'.

8

Select the option of only retrieving proteins with 7 or 8 transmembrane domains by setting the minimum value to 7 and maximum value to 8.

The screenshot shows the GeneDB search interface with the query 'Protein containing one or more predicted transmembrane domains.' selected. Below this, there are two input fields for 'Minimum number of transmembrane domains' and 'Maximum number of transmembrane domains'. The minimum value is set to 7 and the maximum value is set to 8. To the right of each field is a red question mark icon. Below these fields, there is a 'Query options' section with a 'Rows per page' dropdown set to 20 and a red question mark icon. At the bottom, there is a 'Run query' section with 'Submit' and 'Reset' buttons. An arrow points from the instruction box to the 'Submit' button.

9

Press 'Submit'

10

This is the results list of all the proteins containing 7 or 8 predicted transmembrane domains.

[1-20][21-39]					[1-20][21-39]	
	name	id	organism	description		
1		<a href="#">Tb027.1.3380</a>	<i>T. brucei</i>	hypothetical protein, conserved		Manual Annotation
2		<a href="#">Tb04.5020.380</a>	<i>T. brucei</i>	hypothetical protein		Manual Annotation
3		<a href="#">Tb04.24.1.1069</a>	<i>T. brucei</i>	lipophosphoglycan 2, putative		Manual Annotation
4	TbMRPA	<a href="#">Tb04.3112.980</a>	<i>T. brucei</i>	multidrug resistance protein E		Manual Annotation
5		<a href="#">Tb09.160.1010</a>	<i>T. brucei</i>	hypothetical protein		Manual Annotation
6		<a href="#">Tb09.211.0140</a>	<i>T. brucei</i>	chaperone, putative		Manual Annotation
7		<a href="#">Tb09.211.1820</a>	<i>T. brucei</i>	hypothetical protein		Manual Annotation
8		<a href="#">Tb09.244.2570</a>	<i>T. brucei</i>	calcium motive p-type ATPase, putative		Manual Annotation
9		<a href="#">Tb03.4883.650</a>	<i>T. brucei</i>	signal peptide peptidase, putative		Manual Annotation
10		<a href="#">Tb10.61.0820</a>	<i>T. brucei</i>	hypothetical protein, conserved		Manual Annotation
11		<a href="#">Tb10.6k15.0240</a>	<i>T. brucei</i>	hypothetical protein		Manual Annotation
12		<a href="#">Tb10.6k15.1570</a>	<i>T. brucei</i>	aminocalcophosphotransferase, putative		Manual Annotation
13		<a href="#">Tb10.6k15.2020</a>	<i>T. brucei</i>	glucose transporter, fragment		Manual Annotation
14		<a href="#">Tb10.6k15.2450</a>	<i>T. brucei</i>	hypothetical protein		Manual Annotation
15		<a href="#">Tb10.70.0260</a>	<i>T. brucei</i>	mannosyltransferase, putative		Manual Annotation
16	GPI10	<a href="#">Tb10.70.1440</a>	<i>T. brucei</i>	GPI anchor biosynthesis protein		Manual Annotation
17		<a href="#">Tb10.70.3810</a>	<i>T. brucei</i>	hypothetical protein, conserved		Manual Annotation
18		<a href="#">Tb10.70.7920</a>	<i>T. brucei</i>	hypothetical protein, conserved		Manual Annotation
19		<a href="#">Tb10.389.0140</a>	<i>T. brucei</i>	hypothetical protein, conserved		Manual Annotation
20		<a href="#">Tb10.389.0200</a>	<i>T. brucei</i>	hypothetical protein, conserved		Manual Annotation

To view this result set later, or to combine with others [visit the history page](#).  
[Start a completely new complex query](#)

11

You will have now carried out 2 searches, querying GeneDB independently for the predicted ABC transporters as well as for all the proteins with 7 or 8 predicted transmembrane domains. To now identify the subset of ABC transporters with only 7 or 8 transmembrane domains, go to the history page which tracks all the queries you have executed in that session.

These are the descriptions of the queries you have executed

**Query History**

Go To: GeneDB Search: Simple [Help](#)

This page displays results of queries executed using the [boolean](#) query pages. The page will remain empty until searches have been run. Please note that a search combining boolean operators will return multiple results files. Chosen queries are first executed across all organisms within GeneDB and subsequently narrowed down to only return data from the organisms of choice. Once query results have been retrieved, users can combine results files in one of three ways: adding files together (union), identifying common results between files (intersect), or identifying unique results between files (subtraction). **Please note:** you must have cookies enabled in your browser in order for this page to work correctly.

The result files can either be viewed or downloaded as a FASTA file of DNA or protein sequences.

Query	Start time	Response time	Result	Download	Size
<input type="checkbox"/> (Genes for <i>tpp</i> ) <b>intersect</b> (Proteins that contain Pfam domain 'ABC transporter')	3:14:40 PM	<1 second	<a href="#">view</a>	<a href="#">Download</a>	22
<input type="checkbox"/> (Genes for <i>tpp</i> ) <b>intersect</b> (Proteins predicted to have between 7 and 8 transmembrane domains.)	3:15:03 PM	<1 second	<a href="#">view</a>	<a href="#">Download</a>	71

Follow one of the links in the table above to view a query result set or select two of the result sets and use one of the following buttons to combine them. "Union" will create a result set that contains all the genes in either of the selected sets. "Intersect" will create a result set that contains only the genes in both of the selected sets. "Subtract" will remove any genes in the second (i.e., appearing lower in the list) set from the first.

or  or  the selected query results (a new entry will appear at the end of the list)

12

Via the query history page, you can view and download the sequences of your queries for further examination/manipulation. You can also combine results files and/or identify share/unique results between 2 queries. In order to identify the subset of ABC transporters, select both your queries and select 'Intersect' button.

13

This will return a third set of query results which should only contain the subset of ABC transporters.

**GeneDB Query History**

Go To: GeneDB Search: Simple Help

This page displays results of queries executed using the [hoogle](#) query pages. The page will remain empty until searches have been run. Please note that a search combining boolean operators will return multiple results files. Chosen queries are first executed across all organisms within GeneDB and subsequently narrowed down to only return data from the organisms of choice. Once query results have been retrieved, users can combine results files in one of three ways: adding files together (union), identifying common results between files (intersect), or identifying unique results between files (subtraction). **Please note:** you must have cookies enabled in your browser in order for this page to work correctly.

The result files can either be viewed or downloaded as a FASTA file of DNA or protein sequences.

Query	Start time	Response time	Result Download Size
<input type="checkbox"/> ((Genes for <i>tryp</i> ) intersect (Proteins that contain Pfam domain 'ABC transporter'))	3:14:40 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 22
<input type="checkbox"/> ((Genes for <i>tryp</i> ) intersect (Proteins predicted to have between 7 and 8 transmembrane domains.))	3:15:03 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 71
<input type="checkbox"/> (((Genes for <i>tryp</i> ) intersect (Proteins that contain Pfam domain 'ABC transporter')) intersect ((Genes for <i>tryp</i> ) intersect (Proteins predicted to have between 7 and 8 transmembrane domains.)))	1:00:00 AM	<1 second	<a href="#">view</a> <a href="#">Download</a> 2

Follow one of the links in the table above to view a query result set or select two of the result sets and use one of the following buttons to combine them. "Union" will create a result set that contains all the genes in either of the selected sets. "Intersect" will create a result set that contains only the genes in both of the selected sets. "Subtract" will remove any genes in the second (i.e., appearing lower in the list) set from the first.

or  or  the selected query results (a new entry will appear at the end of the list)

14

Click on 'view' to have a look at the results. Clicking on the hyperlinked gene names will take you to the feature pages for these putative ABC transporters.

**GeneDB Results 1 - 2 of 2**

Go To: GeneDB Search: Simple Help

Query: (((Genes for *tryp*) intersect (Proteins that contain Pfam domain 'ABC transporter')) intersect ((Genes for *tryp*) intersect (Proteins predicted to have between 7 and 8 transmembrane domains.)))

[1-2]	name	id	organism	description	[1-2]
1	MRPE, PGPA	<a href="#">Tb927.4.4490</a>	<i>T. brucei</i>	multidrug resistance protein E,P-glycoprotein	Manual annotation
2		<a href="#">Tb09.160.4600</a>	<i>T. brucei</i>	ABC transporter, putative	Manual annotation

To view this result set later, or to download this data [visit the history page](#).

[Start a completely new complex query](#)

You would have been able to retrieve the same result by combining the two queries from the outset using 'AND'. By clicking onto 'AND' on the initial *T. brucei* query page, you will get the option of executing multiple queries simultaneously (see below).

**GeneDB**

Go To: GeneDB Search: Simple Help

This page allows you to build more complex queries against the database using a preset selection of query forms and the boolean operators "AND" and "OR". For example, to construct a simple query begin by clicking on the "AND" button. Then use the pull-down menu to select two different queries. Finally, click on "proceed to next step" to generate a page that will allow you to specify parameters for the two queries. Finishing the resulting boolean query will return only those objects that satisfy query 1 AND query 2. **Please note:** you must select a query from each pull-down menu before the "proceed" button will work. If you need to back up a step, use the browser's "back" button.

You are currently searching *T. brucei* [\(Choose a different organism\)](#)

**AND**

Choose a boolean condition or select a query

OR Proteins containing a specific Pfam domain

Choose a boolean condition or select a query

OR None

Proceed to: Proteins with a product containing a particular keyword or phrase  
Proteins with annotation matching a particular keyword  
Proteins with a predicted GO function  
Proteins with a predicted GO process  
Proteins with a predicted GO component  
Proteins within a range of length in amino acids  
Proteins within a particular range of molecular masses  
Proteins containing a specific Pfam domain  
Proteins with a Pfam domain  
Proteins containing a predicted signal peptide  
Proteins containing one or more predicted transmembrane domains

Hosted by the [Sanger Institute](#)

[Send us your comments on GeneDB](#)

**GeneDB Results 1 - 2 of 2**

Go To: GeneDB Search: Simple Help

Query: (((Genes for *tryp*) intersect (Proteins that contain Pfam domain 'ABC transporter')) intersect ((Genes for *tryp*) intersect (Proteins predicted to have between 7 and 8 transmembrane domains.)))

[1-2]	name	id	organism	description	[1-2]
1	MRPE, PGPA	<a href="#">Tb927.4.4490</a>	<i>T. brucei</i>	multidrug resistance protein E,P-glycoprotein	Manual annotation
2		<a href="#">Tb09.160.4600</a>	<i>T. brucei</i>	ABC transporter, putative	Manual annotation

To view this result set later, or to download this data [visit the history page](#).

[Start a completely new complex query](#)

**Exercise 5    Identifying autotransporters encoded in the genomes of *Bordetella pertussis*, *B. parapertussis* and *B. bronchiseptica*.**

*B. pertussis*, *B. parapertussis* and *B. bronchiseptica* are closely related Gram-negative  $\beta$ -proteobacteria. They colonize the respiratory tract of mammals, causing whooping cough (*B. pertussis*, *B. parapertussis*) as well as a chronic respiratory infection in a range of mammals (*B. bronchiseptica*).

This exercise is designed to identify autotransporters in the three genomes of the *Bordetella* spp. Autotransporters are members of a large family of exported proteins, encoding an integral outer-membrane pore which enables the bacteria to cross the outer membrane. As such, autotransporters are postulated to function in host interaction and virulence, some of which have been experimentally confirmed.

Imagine you came across a recent paper describing the autotransporter complement in *B. bronchiseptica* (see table on page 56).

Now think of ways you could identify autotransporters in the other *Bordetella* species. You could do this in a variety of ways:

- keyword searches of assigned product names (**exercise 4.1**).
- using the orthologue links provided on the gene pages (**exercise 4.2**).
- using BLAST (**exercise 4.3**).
- using Pfam/Reily browsable catalogues (**exercise 4.4**).
- using boolean querying tool (**exercise 4.5**)

Once you've identified the autotransporters across the three species, we're going to examine the genomic loci of one of these transporters a little closer using ACT (**exercise 4.6**).



**Exercise 5.1/5.2**

On the left is the table of putative autotransporters annotated in the *B. bronchiseptica* genome. You could now take a variety of routes to identify autotransporters in the other two genomes. We're going to start using a simple keyword search.

Start by going to the GeneDB homepage at [www.genedb.org](http://www.genedb.org) and choose to go to the *B. paraptussis* homepage by selecting this organism from the pull-down menu. This will get you to the species homepage, providing access to the data via tools (BLAST servers), browsable catalogues and simple search facilities. Type 'autotransporter' into the search box, ensuring that the wild card box is ticked.

**Table 2** Autotransporters encoded in the genomes of *B. pertussis*, *B. paraptussis* and *B. bronchiseptica*

	<i>B. pertussis</i>	<i>B. paraptussis</i>	<i>B. bronchiseptica</i>
SphB1			BB0419
Novel			BB0450
Novel			BB0452
Novel			BB0821
Novel			BB0916
Serum-resistance protein			<u>BB0961</u>
Pertactin			BB1366
BapA/AidB			BB1649
Vag8			BB1864
BapC			BB2033
Phg			BB2246
Novel			BB2270
SphB3 (serine protease)			BB2301
Novel			BB2324
SphB2			BB2741
Novel			BB2830
Novel			BB2941
Novel			BB3110
Novel			BB3111
TcfA			BB3291
BapB			BB3292

Pseudogenes are underlined.

Parkhill *et al.*, Nature Genetics (2003), 35: 32-40.

### *Bordetella paraptussis* GeneDB



Access to the annotation and sequence of *B. paraptussis* strain 12822. This is the result of a collaboration of the Sanger Institute, Julian Parkhill and Andrew Preston of the Centre for Veterinary Science, Dept. of Clinical Veterinary medicine, The University of Bristol.

Parkhill *et al.* (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella paraptussis* and *Bordetella bronchiseptica*. *Nature Genetics* DOI: 10.1038/NG1227. ([PDF version of this article](#))

Database Entry Point

Go To Organisms

Search for gene by ID/description

autotransporter

☒ Include description  
☒ Add wildcards

Full Content Search

Searches/Analysis

[omniBLAST](#)  
[BLAST](#)  
[Motif Search](#)  
[EMOWSE](#)  
[List Download](#)  
[Cross-Organism Search Page](#)  
[Complex/Boolean Query](#)

Browse Catalogues

[Riley](#)  
[Products](#)  
[Pfam](#)

Information

Julian Parkhill

Project manager

Mohammed Sebaiha

Curation

Contact [the developers](#)

Miscellaneous Information

Example genes

[BPP2409](#), [rpsR](#), [rplI](#)

This returns a list of all genes with annotated product lines matching this search term.

GeneDB

## Gene Results List

The Wellcome Trust  
Sanger Institute  
Pathogen Sequencing Unit

Go To

Organisms

Go To

Shortcuts

[Help](#)

Results 1 to 17 of 17 results shown

Previous

Next

Report Download

<i>B. paraptentussis</i>	CDS	<a href="#">BPP2415</a>	autotransporter, vag8
<i>B. paraptentussis</i>	CDS	<a href="#">BPP0735</a>	autotransporter
<i>B. paraptentussis</i>	CDS	<a href="#">BPP0417</a>	autotransporter subtilisin-like protease, sphB1
<i>B. paraptentussis</i>	CDS	<a href="#">BPP2975</a>	putative autotransporter
<i>B. paraptentussis</i>	CDS	<a href="#">BPP2745</a>	autotransporter (pseudogene), sphB2
<i>B. paraptentussis</i>	CDS	<a href="#">BPP0452</a>	autotransporter
<i>B. paraptentussis</i>	CDS	<a href="#">BPP2591</a>	putative autotransporter, bapC
<i>B. paraptentussis</i>	CDS	<a href="#">BPP1815</a>	autotransporter, bapB
<i>B. paraptentussis</i>	CDS	<a href="#">BPP2251</a>	putative autotransporter (Pseudogene), bapA
<i>B. paraptentussis</i>	CDS	<a href="#">BPP0822</a>	autotransporter
<i>B. paraptentussis</i>	CDS	<a href="#">BPP1256</a>	autotransporter
<i>B. paraptentussis</i>	CDS	<a href="#">BPP1618</a>	autotransporter
<i>B. paraptentussis</i>	CDS	<a href="#">BPP1617</a>	autotransporter
<i>B. paraptentussis</i>	CDS	<a href="#">BPP1998</a>	autotransporter, phg
<i>B. paraptentussis</i>	CDS	<a href="#">BPP0449</a>	autotransporter
<i>B. paraptentussis</i>	CDS	<a href="#">BPP2022</a>	autotransporter
<i>B. paraptentussis</i>	CDS	<a href="#">BPP2678</a>	autotransporter (pseudogene)

Click on 'BPP0417' which will take you to the feature page, detailing information associated with this gene.

**Navigation bar pull down menus:**  
You can navigate between different organism datasets and search tools using pull down menus

**Gene name and product information:** The description lines are standardized and indexed so that features sharing the same description lines can be retrieved. Access to the nucleotide and amino acid sequences of the feature are also provided.

**Basic location information and context map:**  
Clicking on the 'Graphical display in Artemis' open up an Artemis applet – which will be discussed further in exercise 2. Via the applet, the feature can be viewed in the context of the sequence and additional annotation.

GeneDB

CDS: BPP0417

Search for:  Go To: Organisms Go To: Shortcuts [Help](#)

**General Information** [Add to Basket](#) [View Basket](#)

Name	BPP0417
Systematic Name	sphB1
Gene Synonyms	
Product	autotransporter subtilisin-like protease
Type	CDS
Sequence	<a href="#">DNA</a> and <a href="#">Protein</a>

**Location**

Chromosome	1
Contig	BPP
Location	complement(434671..437466) Length: 2796 bp
Exons	complement(434671..437466) (Spliced length: 2796 bp)

Context Map: [Graphical Display \(in Artemis\)](#)

**Primary Annotation**

Predicted Peptide Properties			
Mass	99.5 kDa	Amino acids	931
Isoelectric point	pH 9.9	Charge	29.5
Signal Peptide	Not found		
Transmembrane Domains	0 found		

Protein Map:

**Domain Information**

DB	Access	Description	Note
Pfam	<a href="#">PF00082</a>	Subtilase family	HMM/Pfam hit to PF00082, Subtilase family, score 1.9e-10

**Orthologues**

DB	Access	Description	Note
GeneDB_Ebronchireptis	<a href="#">EB0419</a>	autotransporter subtilisin-like protease	
GeneDB_Epertussis	<a href="#">BPP0216</a>	autotransporter subtilisin-like protease	

You'll find manually curated orthologue links towards the bottom of the feature page which will take you to the *B. pertussis* and *B. bronchiseptica* genes.

GeneDB CDS: BP0216

Search for:  Go To: Organisms Shortcuts [Help](#) [Contact us](#)

**General Information** [Add to Basket](#) [View Basket](#)

Name: BP0216  
Systematic Name: *spA1*  
Gene Synonyms: autotransporter subfamily protease  
Product: CDS  
Type: CDS  
Sequence: [DNA](#) [and Protein](#)

**Location**

Chromosome: 1  
Contig: BP  
Location: 223401..226399 Length: 2999 bp  
Exons: 223401..226399 (Spliced length: 2799 bp)

[Graphical Display \(in Artemis\)](#)

Context Map:

[BP0207](#) [BP0208](#) [BP0209](#) [BP0210](#) [BP0211](#) [BP0212](#) [BP0213](#) [BP0214](#) [BP0215](#) [BP0216](#) [BP0217](#) [BP0218](#) [BP0219](#) [BP0220](#) [BP0221](#) [BP0222](#) [BP0223](#) [BP0224](#) [BP0225](#) [BP0226](#) [BP0227](#) [BP0228](#) [BP0229](#) [BP0230](#)

**Primary Annotation**

Previously sequenced as *Bordetella pertussis* autotransporter subfamily protease *spA1* TR:CA044081 (EMBL:AJ131229) (1039 aa) data score: E0 0.100/000% id as 932 aa, and similar to *Xylella fastidiosa* sensor protease XP0126 TR:Q99232 (EMBL:AA003939) (905 aa) data score: E0 1.1e-45, 31.919% id as 943 aa

**Predicted Peptide Properties**

Mass	99.4 kDa	Amino acids	932
Isoelectric point	pI 10.3	Charge	27.0
Signal Peptide	Not found		
Transmembrane Domains	0 found		

Protein Map:

**Domain Information**

DB	Access	Description	Note
Pfam	<a href="#">PF00082</a>	Subfamily family	EMBL:AJ131229 hit to PF00082, Subfamily family
Pfam	<a href="#">PF00082</a>	Subfamily family	EMBL:AJ131229 hit to PF00082, Subfamily family
PROSITE	<a href="#">PS00095</a>	C-5 cytosine-specific DNA methylase C-terminal signature	Scalag:BP hit to PS00095, C-5 cytosine-specific DNA methylase C-terminal signature
PROSITE	<a href="#">PS00138</a>	Sensor proteases, subfamily, sensor active site	Scalag:BP hit to PS00138, Sensor proteases, subfamily, sensor active site

**Orthologues**

DB	Access	Description	Note
GeneDB_Bordetella	<a href="#">BB0419</a>	autotransporter subfamily-like protease	
GeneDB_Bordetella	<a href="#">BB0417</a>	autotransporter subfamily-like protease	

GeneDB CDS: BB0419

Search for:  Go To: Organisms Shortcuts [Help](#) [Contact us](#)

**General Information** [Add to Basket](#) [View Basket](#)

Name: BB0419  
Systematic Name: *spA1*  
Gene Synonyms: autotransporter subfamily protease  
Product: CDS  
Type: CDS  
Sequence: [DNA](#) [and Protein](#)

**Location**

Chromosome: 1  
Contig: BB  
Location: complement(42924..43261) Length: 2796 bp  
Exons: complement(42924..43261) (Spliced length: 2796 bp)

[Graphical Display \(in Artemis\)](#)

Context Map:

[BB0415](#) [BB0416](#) [BB0417](#) [BB0418](#) [BB0419](#) [BB0420](#) [BB0421](#) [BB0422](#) [BB0423](#) [BB0424](#) [BB0425](#) [BB0426](#) [BB0427](#) [BB0428](#) [BB0429](#) [BB0430](#) [BB0431](#) [BB0432](#) [BB0433](#) [BB0434](#) [BB0435](#) [BB0436](#) [BB0437](#) [BB0438](#) [BB0439](#) [BB0440](#) [BB0441](#) [BB0442](#) [BB0443](#) [BB0444](#) [BB0445](#) [BB0446](#) [BB0447](#) [BB0448](#) [BB0449](#) [BB0450](#) [BB0451](#) [BB0452](#) [BB0453](#) [BB0454](#) [BB0455](#) [BB0456](#) [BB0457](#) [BB0458](#) [BB0459](#) [BB0460](#) [BB0461](#) [BB0462](#) [BB0463](#) [BB0464](#) [BB0465](#) [BB0466](#) [BB0467](#) [BB0468](#) [BB0469](#) [BB0470](#) [BB0471](#) [BB0472](#) [BB0473](#) [BB0474](#) [BB0475](#) [BB0476](#) [BB0477](#) [BB0478](#) [BB0479](#) [BB0480](#) [BB0481](#) [BB0482](#) [BB0483](#) [BB0484](#) [BB0485](#) [BB0486](#) [BB0487](#) [BB0488](#) [BB0489](#) [BB0490](#) [BB0491](#) [BB0492](#) [BB0493](#) [BB0494](#) [BB0495](#) [BB0496](#) [BB0497](#) [BB0498](#) [BB0499](#) [BB0500](#) [BB0501](#) [BB0502](#) [BB0503](#) [BB0504](#) [BB0505](#) [BB0506](#) [BB0507](#) [BB0508](#) [BB0509](#) [BB0510](#) [BB0511](#) [BB0512](#) [BB0513](#) [BB0514](#) [BB0515](#) [BB0516](#) [BB0517](#) [BB0518](#) [BB0519](#) [BB0520](#) [BB0521](#) [BB0522](#) [BB0523](#) [BB0524](#) [BB0525](#) [BB0526](#) [BB0527](#) [BB0528](#) [BB0529](#) [BB0530](#) [BB0531](#) [BB0532](#) [BB0533](#) [BB0534](#) [BB0535](#) [BB0536](#) [BB0537](#) [BB0538](#) [BB0539](#) [BB0540](#) [BB0541](#) [BB0542](#) [BB0543](#) [BB0544](#) [BB0545](#) [BB0546](#) [BB0547](#) [BB0548](#) [BB0549](#) [BB0550](#) [BB0551](#) [BB0552](#) [BB0553](#) [BB0554](#) [BB0555](#) [BB0556](#) [BB0557](#) [BB0558](#) [BB0559](#) [BB0560](#) [BB0561](#) [BB0562](#) [BB0563](#) [BB0564](#) [BB0565](#) [BB0566](#) [BB0567](#) [BB0568](#) [BB0569](#) [BB0570](#) [BB0571](#) [BB0572](#) [BB0573](#) [BB0574](#) [BB0575](#) [BB0576](#) [BB0577](#) [BB0578](#) [BB0579](#) [BB0580](#) [BB0581](#) [BB0582](#) [BB0583](#) [BB0584](#) [BB0585](#) [BB0586](#) [BB0587](#) [BB0588](#) [BB0589](#) [BB0590](#) [BB0591](#) [BB0592](#) [BB0593](#) [BB0594](#) [BB0595](#) [BB0596](#) [BB0597](#) [BB0598](#) [BB0599](#) [BB0600](#) [BB0601](#) [BB0602](#) [BB0603](#) [BB0604](#) [BB0605](#) [BB0606](#) [BB0607](#) [BB0608](#) [BB0609](#) [BB0610](#) [BB0611](#) [BB0612](#) [BB0613](#) [BB0614](#) [BB0615](#) [BB0616](#) [BB0617](#) [BB0618](#) [BB0619](#) [BB0620](#) [BB0621](#) [BB0622](#) [BB0623](#) [BB0624](#) [BB0625](#) [BB0626](#) [BB0627](#) [BB0628](#) [BB0629](#) [BB0630](#) [BB0631](#) [BB0632](#) [BB0633](#) [BB0634](#) [BB0635](#) [BB0636](#) [BB0637](#) [BB0638](#) [BB0639](#) [BB0640](#) [BB0641](#) [BB0642](#) [BB0643](#) [BB0644](#) [BB0645](#) [BB0646](#) [BB0647](#) [BB0648](#) [BB0649](#) [BB0650](#) [BB0651](#) [BB0652](#) [BB0653](#) [BB0654](#) [BB0655](#) [BB0656](#) [BB0657](#) [BB0658](#) [BB0659](#) [BB0660](#) [BB0661](#) [BB0662](#) [BB0663](#) [BB0664](#) [BB0665](#) [BB0666](#) [BB0667](#) [BB0668](#) [BB0669](#) [BB0670](#) [BB0671](#) [BB0672](#) [BB0673](#) [BB0674](#) [BB0675](#) [BB0676](#) [BB0677](#) [BB0678](#) [BB0679](#) [BB0680](#) [BB0681](#) [BB0682](#) [BB0683](#) [BB0684](#) [BB0685](#) [BB0686](#) [BB0687](#) [BB0688](#) [BB0689](#) [BB0690](#) [BB0691](#) [BB0692](#) [BB0693](#) [BB0694](#) [BB0695](#) [BB0696](#) [BB0697](#) [BB0698](#) [BB0699](#) [BB0700](#) [BB0701](#) [BB0702](#) [BB0703](#) [BB0704](#) [BB0705](#) [BB0706](#) [BB0707](#) [BB0708](#) [BB0709](#) [BB0710](#) [BB0711](#) [BB0712](#) [BB0713](#) [BB0714](#) [BB0715](#) [BB0716](#) [BB0717](#) [BB0718](#) [BB0719](#) [BB0720](#) [BB0721](#) [BB0722](#) [BB0723](#) [BB0724](#) [BB0725](#) [BB0726](#) [BB0727](#) [BB0728](#) [BB0729](#) [BB0730](#) [BB0731](#) [BB0732](#) [BB0733](#) [BB0734](#) [BB0735](#) [BB0736](#) [BB0737](#) [BB0738](#) [BB0739](#) [BB0740](#) [BB0741](#) [BB0742](#) [BB0743](#) [BB0744](#) [BB0745](#) [BB0746](#) [BB0747](#) [BB0748](#) [BB0749](#) [BB0750](#) [BB0751](#) [BB0752](#) [BB0753](#) [BB0754](#) [BB0755](#) [BB0756](#) [BB0757](#) [BB0758](#) [BB0759](#) [BB0760](#) [BB0761](#) [BB0762](#) [BB0763](#) [BB0764](#) [BB0765](#) [BB0766](#) [BB0767](#) [BB0768](#) [BB0769](#) [BB0770](#) [BB0771](#) [BB0772](#) [BB0773](#) [BB0774](#) [BB0775](#) [BB0776](#) [BB0777](#) [BB0778](#) [BB0779](#) [BB0780](#) [BB0781](#) [BB0782](#) [BB0783](#) [BB0784](#) [BB0785](#) [BB0786](#) [BB0787](#) [BB0788](#) [BB0789](#) [BB0790](#) [BB0791](#) [BB0792](#) [BB0793](#) [BB0794](#) [BB0795](#) [BB0796](#) [BB0797](#) [BB0798](#) [BB0799](#) [BB0800](#) [BB0801](#) [BB0802](#) [BB0803](#) [BB0804](#) [BB0805](#) [BB0806](#) [BB0807](#) [BB0808](#) [BB0809](#) [BB0810](#) [BB0811](#) [BB0812](#) [BB0813](#) [BB0814](#) [BB0815](#) [BB0816](#) [BB0817](#) [BB0818](#) [BB0819](#) [BB0820](#) [BB0821](#) [BB0822](#) [BB0823](#) [BB0824](#) [BB0825](#) [BB0826](#) [BB0827](#) [BB0828](#) [BB0829](#) [BB0830](#) [BB0831](#) [BB0832](#) [BB0833](#) [BB0834](#) [BB0835](#) [BB0836](#) [BB0837](#) [BB0838](#) [BB0839](#) [BB0840](#) [BB0841](#) [BB0842](#) [BB0843](#) [BB0844](#) [BB0845](#) [BB0846](#) [BB0847](#) [BB0848](#) [BB0849](#) [BB0850](#) [BB0851](#) [BB0852](#) [BB0853](#) [BB0854](#) [BB0855](#) [BB0856](#) [BB0857](#) [BB0858](#) [BB0859](#) [BB0860](#) [BB0861](#) [BB0862](#) [BB0863](#) [BB0864](#) [BB0865](#) [BB0866](#) [BB0867](#) [BB0868](#) [BB0869](#) [BB0870](#) [BB0871](#) [BB0872](#) [BB0873](#) [BB0874](#) [BB0875](#) [BB0876](#) [BB0877](#) [BB0878](#) [BB0879](#) [BB0880](#) [BB0881](#) [BB0882](#) [BB0883](#) [BB0884](#) [BB0885](#) [BB0886](#) [BB0887](#) [BB0888](#) [BB0889](#) [BB0890](#) [BB0891](#) [BB0892](#) [BB0893](#) [BB0894](#) [BB0895](#) [BB0896](#) [BB0897](#) [BB0898](#) [BB0899](#) [BB0900](#) [BB0901](#) [BB0902](#) [BB0903](#) [BB0904](#) [BB0905](#) [BB0906](#) [BB0907](#) [BB0908](#) [BB0909](#) [BB0910](#) [BB0911](#) [BB0912](#) [BB0913](#) [BB0914](#) [BB0915](#) [BB0916](#) [BB0917](#) [BB0918](#) [BB0919](#) [BB0920](#) [BB0921](#) [BB0922](#) [BB0923](#) [BB0924](#) [BB0925](#) [BB0926](#) [BB0927](#) [BB0928](#) [BB0929](#) [BB0930](#) [BB0931](#) [BB0932](#) [BB0933](#) [BB0934](#) [BB0935](#) [BB0936](#) [BB0937](#) [BB0938](#) [BB0939](#) [BB0940](#) [BB0941](#) [BB0942](#) [BB0943](#) [BB0944](#) [BB0945](#) [BB0946](#) [BB0947](#) [BB0948](#) [BB0949](#) [BB0950](#) [BB0951](#) [BB0952](#) [BB0953](#) [BB0954](#) [BB0955](#) [BB0956](#) [BB0957](#) [BB0958](#) [BB0959](#) [BB0960](#) [BB0961](#) [BB0962](#) [BB0963](#) [BB0964](#) [BB0965](#) [BB0966](#) [BB0967](#) [BB0968](#) [BB0969](#) [BB0970](#) [BB0971](#) [BB0972](#) [BB0973](#) [BB0974](#) [BB0975](#) [BB0976](#) [BB0977](#) [BB0978](#) [BB0979](#) [BB0980](#) [BB0981](#) [BB0982](#) [BB0983](#) [BB0984](#) [BB0985](#) [BB0986](#) [BB0987](#) [BB0988](#) [BB0989](#) [BB0990](#) [BB0991](#) [BB0992](#) [BB0993](#) [BB0994](#) [BB0995](#) [BB0996](#) [BB0997](#) [BB0998](#) [BB0999](#) [BB1000](#) [BB1001](#) [BB1002](#) [BB1003](#) [BB1004](#) [BB1005](#) [BB1006](#) [BB1007](#) [BB1008](#) [BB1009](#) [BB1010](#) [BB1011](#) [BB1012](#) [BB1013](#) [BB1014](#) [BB1015](#) [BB1016](#) [BB1017](#) [BB1018](#) [BB1019](#) [BB1020](#) [BB1021](#) [BB1022](#) [BB1023](#) [BB1024](#) [BB1025](#) [BB1026](#) [BB1027](#) [BB1028](#) [BB1029](#) [BB1030](#) [BB1031](#) [BB1032](#) [BB1033](#) [BB1034](#) [BB1035](#) [BB1036](#) [BB1037](#) [BB1038](#) [BB1039](#) [BB1040](#) [BB1041](#) [BB1042](#) [BB1043](#) [BB1044](#) [BB1045](#) [BB1046](#) [BB1047](#) [BB1048](#) [BB1049](#) [BB1050](#) [BB1051](#) [BB1052](#) [BB1053](#) [BB1054](#) [BB1055](#) [BB1056](#) [BB1057](#) [BB1058](#) [BB1059](#) [BB1060](#) [BB1061](#) [BB1062](#) [BB1063](#) [BB1064](#) [BB1065](#) [BB1066](#) [BB1067](#) [BB1068](#) [BB1069](#) [BB1070](#) [BB1071](#) [BB1072](#) [BB1073](#) [BB1074](#) [BB1075](#) [BB1076](#) [BB1077](#) [BB1078](#) [BB1079](#) [BB1080](#) [BB1081](#) [BB1082](#) [BB1083](#) [BB1084](#) [BB1085](#) [BB1086](#) [BB1087](#) [BB1088](#) [BB1089](#) [BB1090](#) [BB1091](#) [BB1092](#) [BB1093](#) [BB1094](#) [BB1095](#) [BB1096](#) [BB1097](#) [BB1098](#) [BB1099](#) [BB1100](#) [BB1101](#) [BB1102](#) [BB1103](#) [BB1104](#) [BB1105](#) [BB1106](#) [BB1107](#) [BB1108](#) [BB1109](#) [BB1110](#) [BB1111](#) [BB1112](#) [BB1113](#) [BB1114](#) [BB1115](#) [BB1116](#) [BB1117](#) [BB1118](#) [BB1119](#) [BB1120](#) [BB1121](#) [BB1122](#) [BB1123](#) [BB1124](#) [BB1125](#) [BB1126](#) [BB1127](#) [BB1128](#) [BB1129](#) [BB1130](#) [BB1131](#) [BB1132](#) [BB1133](#) [BB1134](#) [BB1135](#) [BB1136](#) [BB1137](#) [BB1138](#) [BB1139](#) [BB1140](#) [BB1141](#) [BB1142](#) [BB1143](#) [BB1144](#) [BB1145](#) [BB1146](#) [BB1147](#) [BB1148](#) [BB1149](#) [BB1150](#) [BB1151](#) [BB1152](#) [BB1153](#) [BB1154](#) [BB1155](#) [BB1156](#) [BB1157](#) [BB1158](#) [BB1159](#) [BB1160](#) [BB1161](#) [BB1162](#) [BB1163](#) [BB1164](#) [BB1165](#) [BB1166](#) [BB1167](#) [BB1168](#) [BB1169](#) [BB1170](#) [BB1171](#) [BB1172](#) [BB1173](#) [BB1174](#) [BB1175](#) [BB1176](#) [BB1177](#) [BB1178](#) [BB1179](#) [BB1180](#) [BB1181](#) [BB1182](#) [BB1183](#) [BB1184](#) [BB1185](#) [BB1186](#) [BB1187](#) [BB1188](#) [BB1189](#) [BB1190](#) [BB1191](#) [BB1192](#) [BB1193](#) [BB1194](#) [BB1195](#) [BB1196](#) [BB1197](#) [BB1198](#) [BB1199](#) [BB1200](#) [BB1201](#) [BB1202](#) [BB1203](#) [BB1204](#) [BB1205](#) [BB1206](#) [BB1207](#) [BB1208](#) [BB1209](#) [BB1210](#) [BB1211](#) [BB1212](#) [BB1213](#) [BB1214](#) [BB1215](#) [BB1216](#) [BB1217](#) [BB1218](#) [BB1219](#) [BB1220](#) [BB1221](#) [BB1222](#) [BB1223](#) [BB1224](#) [BB1225](#) [BB1226](#) [BB1227](#) [BB1228](#) [BB1229](#) [BB1230](#) [BB1231](#) [BB1232](#) [BB1233](#) [BB1234](#) [BB1235](#) [BB1236](#) [BB1237](#) [BB1238](#) [BB1239](#) [BB1240](#) [BB1241](#) [BB1242](#) [BB1243](#) [BB1244](#) [BB1245](#) [BB1246](#) [BB1247](#) [BB1248](#) [BB1249](#) [BB1250](#) [BB1251](#) [BB1252](#) [BB1253](#) [BB1254](#) [BB1255](#) [BB1256](#) [BB1257](#) [BB1258](#) [BB1259](#) [BB1260](#) [BB1261](#) [BB1262](#) [BB1263](#) [BB1264](#) [BB1265](#) [BB1266](#) [BB1267](#) [BB1268](#) [BB1269](#) [BB1270](#) [BB1271](#) [BB1272](#) [BB1273](#) [BB1274](#) [BB1275](#) [BB1276](#) [BB1277](#) [BB1278](#) [BB1279](#) [BB1280](#) [BB1281](#) [BB1282](#) [BB1283](#) [BB1284](#) [BB1](#)

**Exercise 5.3**

Using the BLAST server

Go to the *B. bronchiseptica* feature page for BB1366.

GeneDB

CDS: BB1366

*B. bronchiseptica*  
GeneDB

Search for  Go To  Organisms  Go To  Shortcuts  [Help](#) [Contact curator](#)

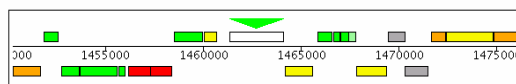
General Information <a href="#">Add to Basket</a> <a href="#">View Basket</a>	
Name	BB1366
Systematic Name	prn
Gene Synonyms	pertactin precursor
Product	CDS
Type	<a href="#">DNA and Protein</a>

Retrieve the amino acid sequence by clicking on the sequence link.

Location	
Chromosome	1
Contig	BB
Location	1461346..1464096 Length: 2751 bp
Exons	1461346..1464096 (Spliced length: 2751 bp)

[Graphical Display \(in Artemis\)](#)

Context Map:



<a href="#">BB1357</a> <a href="#">BB1358</a> <a href="#">BB1359</a> <a href="#">BB1360</a> <a href="#">BB1361</a> <a href="#">BB1362</a> <a href="#">BB1363</a> <a href="#">BB1364</a> <a href="#">BB1365</a> <a href="#">BB1366</a> <a href="#">BB1367</a> <a href="#">BB1368</a> <a href="#">BB1369</a> <a href="#">BB1370</a> <a href="#">BB1371</a> <a href="#">BB1372</a> <a href="#">BB1373</a> <a href="#">BB1374</a> <a href="#">BB1375</a> <a href="#">BB1376</a> <a href="#">BB1377</a>	Primary Annotation	
	Predicted Peptide Properties	
Mass	94.3 kDa	Amino acids
Isoelectric point	pH 10.0	Charge
Signal Peptide	Not found	23.0
Transmembrane Domains	0 found	

Click on 'send to omniBLAST' link. omniBLAST permits searching across different databases selected by the user.

```

-----
GCCAACACCA TGCTGCTGGT GCAGACGCCA CGAGGCAGCG CGGCGACCTT TACCCCTTGC
AACAAAGGACG GCAAGGTCGA TATCGGTACC TACCGCTATC GATTGGCCGC CAACGGCAAT
GGGCAGTGGA GCTGTGGTGG CGCGAAGGCG CGGCCGGCGC CCAAGCCCGC GCCGCGAGCC
GGTCCCCAGC CGGTCCCCA GCGGCCGAG CGGCCGAGC GCGCCAGAG GAGCGCGGAA
GCGCGCGCGC CGCAACCGCC GCGGGGCGAG GAGTTGTCCG CGGCCGCCAA CGCGCGGGTC
AACACGGGTG GGGTGGGGCT GCGGAGCAGC CTCTGGTACG CCGAAAGCAA TGGCTTGTCC
AAGCGCGCTGG GCGAGTTGCG CCTGAATCCG GACGCCGGCG GCGCTTGGGG CCGCGGCTTC
GCGCAACGCG AGCAACTGGA CAACCGCGCC GGGCGGCGCT TCGACCGAGAA GGTGGCGGGC
TTGAGCTGG GCGCCGACCA CGCGGTGGCG GTGGCCGGCG GCGCTGGSCA CCGGGCGGGG
CTGGCCGGCT ATACCGCGCG GAGCCGCGCG TTTACCGCGC ACGCGCGCGC CCACACCGAC
AGCGTGCATG TCGGGGGCTA TGCCACCTAT ATCGCCAAAC GCGGTTTCTA CCGTGACGCG
ACGCTGCGCG CCAGCGCGCT CGAAATGAC TTCAGGTGG CGGCCAGCGA TGGGTACGCG
GTCAAGGGCA AGTACCGCAC CCATGGGGTA GCGGCTTCGC TCGAGCGCGG CCGGCGCTTC
GCCCATGCGC ACGGCTGGTT CCTCGAGCGC GAGGCCGAGC TGGCGGTGTT CCGGGTGGCG
GGCGGTGCGT ACCGCGCGCG CAATGGCTCG GCGGTGCGCG ACGAAGCGCG CAGCTCGGTG
CTGGGTGCGC TGGGCTGGGA GGTGCGGCAAG CGCATCGAAC TGGCAGCGCG CAGGCAGGTG
CAGCCATACA TCAAGGCCAG CGTGTTCAG GAGTTCGAGC GCGCGGGTAC GGTACGACCC
AACGCGATCG CGCACCGCAC CGAATCGCG GGCACGCGCG CCGAAGTGG CCGTGGCATG
GCGCGCGCGC TGGGCGCGCG CCACAGCGCT TATGCTCGT ACGAGTACT CAAGGGCGCG
AAGCTGCCCA TGCCGTGGAC CTTCACCGCG GGCTACCGGT ACGCTGGTA A

```

[Send to GeneDB omniBLAST](#) [Send to GeneDB BLAST](#) [Send to BLAST at NCBI](#)

```

>BB1366 ||prn|pertactin precursor|Bordetella bronchiseptica|chr 1||Manual
MNMSLSRIVK AAPLRRTTLA MALGALGALG AAPAAHADWN NQSIKAGER QHGIHQKSD
GAGVRTATGT TIKVSGRQAG GVLLNPAAE LRFQNGSVTS SGQLFDEGVR RFLGTVTVKA
GKLVADHATL ANVSDTRDDG GIALYVAGEQ AQASIASTL QGAGGVVER GANVTVQST
IKLVGHIGT LQPLQPEDLP PSRVVLGDTL VTAVPASGAP AAUVSVFGANE LTVDDGGHITG
GAAAGVAAMD GAIVHLQRAI IRRGDAPAGG AVPPGAVPVG AVPPGGFPLL DGWGVVDVSD
STVDLAQSIIV EAPQLGAAIR AGRGARVTVS GGSLSAPHGN VIETGGGARR PPPASPLSI
TLQAGARAQG RALLYRVLPF PVKLTLAGGA QGGQDIVATE LPPIPGASSG PLDVALASQA
RNTGATRAVD SLSDNATUV MTDNSNVGAL RLASDGSVDF QQPAAEAGRFK VLMVDTLAAGS
GLFRMNVFAD LGLSDKLVMV RDASGQHLRW VRNSGSEFAS ANTMILLVQTP RGSAAFTFLA
NKDGKVDIGT YRYRLAANGN QWLSLVGAKA PPAPKPAPQP GPQPGPQPPQ PPQPPQRPQE
APAPQPPAPG ELASAAANAIV NTGGVGLAST LWYAESNALS KRLGELRLNP DAGGAWGRGF
AQRQQLDNRA GRRFDQKVAG FELGADHAVA VAGGRWHLGG LAGYTRGDRG FTGDDGGHITD
SVHVGGYATY IANSGFYLDL TLRASLEND FKVAGSDGYA VKGYRTHGV GASLEAGRRF
AHADGUFLEP QAELAVFRVG GGAYRAANGL RVRDEGGSSV LGRLGLEVGK RIELAGGRQV
QPYIKASVLQ EFDGAGTVRT NGIAHRTCLR GTRAEGLGLH AALGRGHSLS YASYEYSKGP
KLAMPUTFHA GTRYSW

```

**GeneDB** **GeneDB omniBLAST Server** 

Go To  Go To  [Help](#)

OmniBLAST will perform a BLAST search on a set of protein databases (BLASTP or BLASTX, depending on the query sequence) or nucleotide databases (BLASTN and TBLASTX or TBLASTN) available in GeneDB and return a list of the best five HSP for each database. If there are any HSP you can click on Full Search to see the complete BLAST output.

To search individual databases with different parameters  single organism BLAST

**QUERY DATA**

Paste your sequence here. [fasta format](#) or just plain text will do

```
>BB1366 |||prn|pertactin precursor|Bordetella
bronchiseptica|chr_11|||B8a001
NNMSLSRIVKALPLRTTLAMALGALGAAPAAHADVNDVOSII
KAGERQGIHLKQSD
GAQVETATOTTIRVSGRQAQGVLLKNPAAELRFPQNSVTSRQGLF
DEGVRFLLGTIVTKA
```

Determine sequence type automatically ☐ or set sequence type to DNA ☐ protein ☒

*Note: OmniBLAST searches may take several minutes depending on the number of selected databases. Please check the databases chosen below are correct*

**DATABASE OPTIONS**

Search *only* the BLAST databases selected below

[Jump down page to: Fungi](#) [Protists](#) [Bacteria](#) [Eukaryotes](#)

**A. fumigatus**

☐ A. fumigatus clustered ESTs ☒ A. fumigatus predicted proteins

**B. bronchiseptica**

☒ B. bronchiseptica complete sequence ☒ B. bronchiseptica predicted proteins

**B. parapertussis**

☐ B. parapertussis complete sequence ☒ B. parapertussis predicted proteins

**B. pertussis**

☐ B. pertussis complete sequence ☒ B. pertussis predicted proteins

**C. diphtheriae**

☐ C. diphtheriae predicted genes (coding sequences) ☐ C. diphtheriae predicted proteins

**B. pseudomallei**

☐ B. pseudomallei predicted genes (coding sequences) ☐ B. pseudomallei predicted proteins

**E. carotovora subsp. atroseptica**

☐ E. carotovora predicted genes (coding sequences) ☐ E. carotovora predicted proteins

**S. aureus subsp. aureus strain MRSA252**

☐ S. aureus MRSA predicted genes (coding sequences) ☐ S. aureus MRSA proteins

**S. aureus subsp. aureus MSSA476**

☐ S. aureus MSSA predicted genes (coding sequences) ☐ S. aureus MSSA proteins

**S. coelicolor**

☐ S. coelicolor predicted genes (coding sequences) ☐ S. coelicolor proteins

**S. typhi**

☐ S. typhi chromosomal sequence ☐ S. typhi chromosomal proteins

☐ S. typhi pHCM1 sequence ☐ S. typhi pHCM1 proteins

☐ S. typhi pHCM2 sequence ☐ S. typhi pHCM2 proteins

**G. moritans**

☐ G. moritans clustered ESTs ☐ G. moritans clustered ESTs translated sequences

**RESETS**

*Note: OmniBLAST searches may take several minutes depending on the number of selected databases*

This is the omniBLAST page, providing access to databases of all sequences housed in GeneDB. By default, the datasets of the organism you started out from will be selected. For this search select the three *Bordetella* spp. protein databases. Note that the amino acid sequence of the protein has automatically been pasted into the query sequence box.

Start omniBLAST by clicking on the ‘Start omniBLAST’ button.

Retrieve the results by clicking on the ‘retrieve’ button. As indicated, results will be accessible for the next 2 weeks using the listed URL.

**GeneDB**

OmniBLAST Server Submission

Go To  Search

Retrieve result for id:

Your BLAST query has been added to the queue of jobs.  
The majority of BLASTs are completed within two minutes.

To retrieve your results, click the **retrieve** button above, or use the following URL: <http://www.genedb.org/genedb2/blast/getblast?id=s2dPAI71A78Cm959999725>



You will retrieve an abbreviated BLAST results page, listing only the top 5 hits without alignments. Click on the '**Full BLAST Search**' of your query sequence against the predicted *B. paraptussis* proteome.

## Blast Server Results

Search Simple [Help](#)

Retrieve result for id:

At peak times your BLAST searches could take longer than normal. Please be patient.

BLAST results are kept on our [servers](#) for three days following query submission. Results may be retrieved any number of times during this period. After this time queries must be resubmitted if further examination is required.

Summary for: *B. paraptussis* predicted proteins [wublastp], for query: BPP0417 [\[Full BLAST Search\]](#)

Name: <a href="#">sphB1</a>	Score: 4845	(P/N): 0.	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">BPP0452</a>	Score: 236	(P/N): 2.2e-17	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">BPP0822</a>	Score: 226	(P/N): 1.8e-16	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">BPP0449</a>	Score: 218	(P/N): 2.4e-15	N: 2	<a href="#">[Full Sequence]</a>
Name: <a href="#">sphB3</a>	Score: 197	(P/N): 1.8e-13	N: 1	<a href="#">[Full Sequence]</a>

Summary for: *B. pertussis* predicted proteins [wublastp], for query: BPP0417 [\[Full BLAST Search\]](#)

Name: <a href="#">sphB1</a>	Score: 4755	(P/N): 0.	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">BPP0529</a>	Score: 215	(P/N): 1.6e-18	N: 2	<a href="#">[Full Sequence]</a>
Name: <a href="#">sphB3</a>	Score: 203	(P/N): 3.3e-14	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">sphB2</a>	Score: 208	(P/N): 6.1e-14	N: 3	<a href="#">[Full Sequence]</a>
Name: <a href="#">BPP0775</a>	Score: 168	(P/N): 2.6e-11	N: 1	<a href="#">[Full Sequence]</a>

Summary for: *B. bronchiseptica* predicted proteins [wublastp], for query: BPP0417 [\[Full BLAST Search\]](#)

Name: <a href="#">sphB1</a>	Score: 4827	(P/N): 0.	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">BB0452</a>	Score: 233	(P/N): 7.8e-17	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">BB0916</a>	Score: 226	(P/N): 2.3e-16	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">sphB2</a>	Score: 223	(P/N): 2.2e-15	N: 2	<a href="#">[Full Sequence]</a>
Name: <a href="#">BB0450</a>	Score: 218	(P/N): 5.5e-15	N: 2	<a href="#">[Full Sequence]</a>

## Blast Server Results

Go To GeneDB Search Simple [Help](#)

Retrieve result for id:

At peak times your BLAST searches could take longer than normal. Please be patient.

BLAST results are kept on our [servers](#) for three days following query submission. Results may be retrieved any number of times during this period. After this time queries must be resubmitted if further examination is required.

Low complexity filtering disabled  
Repeatmasker disabled

BLASTP 2.0.0-MP-WashU [16-Sep-2002] [decunika.0-ev6-132LFF64 2002-09-18T19:28:12]

Copyright (C) 1996-2002 Washington University, Saint Louis, Missouri USA.  
All Rights Reserved.

Reference: Gish, W. (1996-2002) <http://blast.wustl.edu>

Query= BB1366

(516 letters)

Database: BPA.pap

4185 sequences; 1,373,478 total letters.

Searching...10...20...30...40...50...60...70...80...90...100% done

Sequences producing High-scoring Segment Pairs:		High Score	Probability P(N)	Smallest Sum	N
<a href="#">BPP1150</a>	pertactin precursor 1229069:1231837 forward ...	4642	0.	1	
<a href="#">hagC</a>	BPP2591 putative autotransporter 2791262:2794223 fo...	1121	5.4e-161	2	
<a href="#">BPP2925</a>	putative autotransporter 3197029:3199842 reverse ...	1062	2.3e-147	2	
<a href="#">BPP1618</a>	autotransporter 1727612:1729171 forward MW:53634	945	4.5e-100	3	
<a href="#">BPP1617</a>	autotransporter 1727508:1727183 forward MW:52613	1012	1.7e-106	2	
<a href="#">hagB</a>	BPP1815 autotransporter 1930725:1940542 forward MW:...	1015	2.3e-104	1	
<a href="#">BPP2022</a>	autotransporter 2161432:2163048 forward MW:57046	966	1.6e-103	2	
<a href="#">yagB</a>	BPP2415 autotransporter 2587395:2590142 forward MW:...	633	3.5e-82	2	
<a href="#">yagL</a>	BPP1999 autotransporter 2130703:2140030 forward MW:4...	329	9.6e-34	1	
<a href="#">BPP0725</a>	autotransporter 700120:703002 reverse MW:102709	206	6.0e-20	2	
<a href="#">BPP0449</a>	autotransporter 472347:477197 forward MW:159867	329	1.9e-27	1	
<a href="#">BPP0452</a>	autotransporter 478053:483362 forward MW:177085	223	1.1e-20	2	
<a href="#">BPP0822</a>	autotransporter 803503:807405 reverse MW:129686	119	2.1e-08	3	
<a href="#">hagB</a>	BPP3027 filamentous hemagglutinin/adhesin 3256839:3...	124	3.2e-08	2	
<a href="#">sphB3</a>	BPP2053 serine protease 2193178:2196408 forward MW:...	121	7.1e-07	2	
<a href="#">BPP3791</a>	Proline-rich inner membrane protein 4104773:41058...	126	6.7e-06	2	
<a href="#">BPP2368</a>	putative membrane protein 2530013:2532498 reverse...	116	8.1e-06	2	
<a href="#">yagL</a>	BPP2489 adhesin 2468735:2481391 reverse MW:439845	111	8.2e-06	2	
<a href="#">yagS</a>	BPP1243 adhesin 1331868:1339673 forward MW:268245	110	0.00017	3	
<a href="#">BPP0732</a>	hypothetical protein 776060:777270 forward MW:40073	113	0.00017	2	
<a href="#">BPP1520</a>	conserved hypothetical protein 1626284:1627114 fo...	103	0.00032	1	
<a href="#">tonB</a>	BPP2535 siderophore-mediated iron transport protein...	102	0.00042	1	
<a href="#">fliX</a>	BPP0849 cell division protein 918936:920003 forward...	103	0.00051	1	
<a href="#">fliX</a>	<a href="#">fliA</a> , <a href="#">fliA</a> , <a href="#">fliA</a> , BPP1505 flagellar hook-length control p...	80	0.00082	2	
<a href="#">dnaX</a>	<a href="#">dnaX</a> , <a href="#">dnaX</a> , BPP1221 DNA polymerase III subunit Tan...	105	0.00046	3	
<a href="#">BPP3783</a>	hypothetical protein 4096537:4098510 forward MW:6...	72	0.0072	2	
<a href="#">BPP3785</a>	putative secreted protein 944579:944785 forward...	81	0.0077	1	

The BLAST result reveals 13 genes in *B. paraptussis* which have a High degree of sequence similarity. Notice they are all annotated as Autotransporters.

Query:	1	MHMSL.SR.VYKAAPLRRTTLAMALGALGALGAAPAAHWDNNQSIIKAERQHGHITHIKQSD	60
		MHMSL.SR.VYKAAPLRRTTLAMALGALGAAPAAHWDNNQSIIKAERQHGHITHIKQSD	
Shjet:	1	MHMSL.SR.VYKAAPLRRTTLAMALGALGAAPAAHWDNNQSIIKAERQHGHITHIKQSD	57
Query:	61	GAGVRTATGTTIKVSGRQAGVLLNPAEELRFQNGSVTSSGQLPDEGVRRLGTVTVKA	120
		GAGVRTATGTTIKVSGRQAGVLLNPAEELRFQNGSVTSSGQLPDEGVRRLGTVTVKA	
Shjet:	58	GAGVRTATGTTIKVSGRQAGVLLNPAEELRFQNGSVTSSGQLPDEGVRRLGTVTVKA	117
Query:	121	GKLVADHATLANVSDTRDDDGIALYVAGEQAQASIASTLTLQAGGVRVERGANVTVQRST	180
		GKLVADHATLANVSDTRDDDGIALYVAGEQAQASIASTLTLQAGGVRVERGANVTVQRST	
Shjet:	118	GKLVADHATLANVSDTRDDDGIALYVAGEQAQASIASTLTLQAGGVRVERGANVTVQRST	177
Query:	181	IVDGGHIGITLQPLQPEDLPSSRVLLGDTSVTAVPASGAPAAVSVFGANELIVDGGHITG	240
		IVDGGHIGITLQPLQPEDLPSSRVLLGDTSVTAVPASGAPAAVSVFGANELIVDGGHITG	
Shjet:	178	IVDGGHIGITLQPLQPEDLPSSRVLLGDTSVTAVPASGAPAAVSVFGANELIVDGGHITG	237
Query:	241	GRAAGVAAMDGAIVHLQRATIRRDAPAGGAVPGGAVPGGAVPGGGLPLDGYGVDS	300
		GRAAGVAAMDGAIVHLQRATIRRDAPAGGAVPGGAVPGGAVPGGGLPLDGYGVDS	
Shjet:	238	GRAAGVAAMDGAIVHLQRATIRRDAPAGGAVPGGAVPGGAVPGGGLPLDGYGVDS	297
Query:	301	STVYDLAGSIVEAPQLGAATIRAGRGARVTVSG6SLAPHGNVITG6GARFFPPASP.LSI	360
		STVYDLAGSIVEAPQLGAATIRAGRGARVTVSG6SLAPHGNVITG6GARFFPPASP.LSI	
Shjet:	298	STVYDLAGSIVEAPQLGAATIRAGRGARVTVSG6SLAPHGNVITG6GARFFPPASP.LSI	357
Query:	361	TLQAGARAQGRALLYRVLPEPVKLTLAGGAQGQGDIVATELPPIGASSGPLDVALASQA	420
		TLQAGARAQGRALLYRVLPEPVKLTLAGGAQGQGDIVATELPPIGASSGPLDVALASQA	
Shjet:	358	TLQAGARAQGRALLYRVLPEPVKLTLAGGAQGQGDIVATELPPIGASSGPLDVALASQA	417
Query:	421	RWTGATRAVDSLSDNATWMTDINSNGALRLASDGVDFQPPAEGRFKVLMDTLA6S	480
		RWTGATRAVDSLSDNATWMTDINSNGALRLASDGVDFQPPAEGRFKVLMDTLA6S	
Shjet:	418	RWTGATRAVDSLSDNATWMTDINSNGALRLASDGVDFQPPAEGRFKVLMDTLA6S	477

Selecting the link for the top BLAST hit will take you to the Multiple alignment of the *B. bronchiseptica* query sequence and the likely ortholog in *B. paraptussis*.

Open a new window and go to the *B. pertussis* homepage and you can access the putative orthologue by typing in the gene name.



This data described in: Parkhill *et al* (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nature Genetics* DOI: 10.1038/Ng1227. (PDF version of this article)

**Database Entry Point**

Search for  by ID/description

☒ Include description  
☒ Add wildcards

Full Content Search

**Go To** Organisms

**Go To** Organisms

**Searches/An**

[omniBLAST](#)

[BLAST](#)

[List Downl](#)

[Cross-Organism S](#)

[Complex/Boolea](#)

**CDS: BP1054**

**GeneDB**

B. pertussis

[Help](#)  
[Contact curator](#)

Search for

**Go To** Organisms

**Go To** Shortcuts

**General Information** [Add to Basket](#) [View Basket](#)

**Go To** Organisms

**Name**

**Systematic Name**

**Gene Synonyms**

**Product**

**Type**

**Sequence**

BP1054

pm

pertactin precursor

CDS

[DNA](#) and [Protein](#)

**Location**

**Chromosome**

**Contig**

**Location**

**Exons**

1

BP

1098091..1100823 Length: 2733 bp

1098091..1100823 (Spliced length: 2733 bp)

**Graphical Display (in Artemis)**

**Context Map:**

**Example genes**

[Phg](#), [BapB](#), [Bfi](#)

**BP1045 BP1046 BP1047 BP1048 BP1049 BP1050 BP1051 BP1052 BP1053 >BP1054< BP1055 BP1056 BP1057 BP1058 BP1059 BP1060**

**BP1061 BP1062 BP1063 BP1064 BP1065**

**Primary Annotation**

Identical to the previously sequenced *Bordetella pertussis* pertactin precursor Pm or Cmp69A SW:PERT\_BORPE (P14283) (910 aa) fasta scores: E(0, 100% id in 910 aa, and to *Bordetella bronchiseptica* pertactin precursor Pm SW:PERT\_BORBR (Q03035) (911 aa) fasta scores: E(0, 14e-160, 91.31% id in 921 aa

**Mass**

**Isoelectric point**

93.4 kDa

pH 10.0

**Amino acids**

**Charge**


910

17.0


**Predicted Peptide Properties**

**Primary Annotation**

## Exercise 5.4 Using browsable catalogues (Riley)



### *Bordetella parapertussis* GeneDB



This page provides access to the annotation and sequence of *B. parapertussis* strain 12822. This is the result of a collaboration of the Sanger Institute with [Duncan Maskell](#) and [Andrew Preston](#) of the Centre for Veterinary Science, [Dept. of Clinical Veterinary medicine](#), The University of Cambridge.

This data described in: Parkhill *et al* (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. [Nature Genetics](#) DOI: 10.1038/Ng1227. ([PDF version of this article](#))

Database Entry Point

Go To Organisms

Search for gene  
by ID/description

☐ Include description  
☒ Add wildcards

Full Content Search

Searches/Analysis

[omniBLAST](#)  
[BLAST](#)  
[List Download](#)  
[Cross-Organism Search Page](#)  
[Complex/Boolean Query](#)

Browse Catalogues

[Riley](#)  
[Products](#)  
[Pfam](#)

Julian Parkhill  
Mohammed Sebaiha

Contact [the developers](#)

Project manager  
Curation

Example genes

[BPP2409](#), [rpsB](#), [rplI](#)

*B. parapertussis* [project page](#) (at the Sanger Institute)  
*B. parapertussis* [genome](#) (by FTP from the Sanger Institute)

GeneDB provides access to browsable catalogues (product lines, Pfam predictions, Riley classification). Click on 'Riley' which will take you to a list of each of the categories linked to genes annotated to this term.

Click on 'Pathogenicity Islands/determinants'

- [4.1.1 inner membrane](#) (208)
- [4.1.2 Murein sacculus, peptidoglycan](#) (35)
- [4.1.3 Outer membrane constituents](#) (26)
- [4.1.4 Surface polysaccharides & antigens](#) (38)
- [4.1.5 Surface structures](#) (9)
- [4.2.2 Ribosomal proteins - synthesis, modification](#) (53)
- [4.2.3 Ribosomes - maturation and modification](#) (3)
- [5.1.1 Colicin-related functions](#) (2)
- [5.1.2 Phage-related functions and prophages](#) (55)
- [5.1.4 Transposon-related functions](#) (260)
- [5.1.5 Pathogenicity Islands/determinants](#) (34)
- [6.1.1 Global regulatory functions](#) (308)
- [7.0.0 Not classified \(included putative assignments\)](#) (386)


This brings up a list of all the genes annotated to this term. Are the autotransporters you're looking for listed here?

[Report Download](#)

<i>B. pertussis</i>	CDS	<a href="#">BP0216</a>	autotransporter subtilisin-like protease, sphB1
<i>B. pertussis</i>	CDS	<a href="#">BP0529</a>	autotransporter
<i>B. pertussis</i>	CDS	<a href="#">BP0758</a>	cyclolysin-activating lysine-acyltransferase, cyaC, HlyC
<i>B. pertussis</i>	CDS	<a href="#">BP0760</a>	bifunctional hemolysin-adenylate cyclase precursor, cyaA, cya
<i>B. pertussis</i>	CDS	<a href="#">BP0761</a>	cyclolysin secretion ATP-binding protein, cyaB
<i>B. pertussis</i>	CDS	<a href="#">BP0762</a>	cyclolysin secretion protein, cyaD
<i>B. pertussis</i>	CDS	<a href="#">BP0763</a>	cyclolysin secretion protein, cyaE
<i>B. pertussis</i>	CDS	<a href="#">BP0874</a>	vir-repressed protein, vir-18
<i>B. pertussis</i>	CDS	<a href="#">BP1054</a>	pertactin precursor, pm
<i>B. pertussis</i>	CDS	<a href="#">BP1110</a>	serine protease, sphB3
<i>B. pertussis</i>	CDS	<a href="#">BP1112</a>	putative outer membrane ligand binding protein, bapA
<i>B. pertussis</i>	CDS	<a href="#">BP1200</a>	autotransporter (pseudogene), bapB
<i>B. pertussis</i>	CDS	<a href="#">BP1201</a>	tracheal colonization factor precursor, tcfA
<i>B. pertussis</i>	CDS	<a href="#">BP1251</a>	putative toxin
<i>B. pertussis</i>	CDS	<a href="#">BP1344</a>	autotransporter
<i>B. pertussis</i>	CDS	<a href="#">BP1610</a>	putative autotransporter (Pseudogene)
<i>B. pertussis</i>	CDS	<a href="#">BP1660</a>	autotransporter, sphB2
<i>B. pertussis</i>	CDS	<a href="#">BP1767</a>	autotransporter, phg
<i>B. pertussis</i>	CDS	<a href="#">BP1793</a>	autotransporter (Pseudogene)
<i>B. pertussis</i>	CDS	<a href="#">BP1884</a>	hemolysin activator-like protein, fhaC
<i>B. pertussis</i>	CDS	<a href="#">BP2224</a>	putative autotransporter, bapA
<i>B. pertussis</i>	CDS	<a href="#">BP2315</a>	autotransporter, vag8
<i>B. pertussis</i>	CDS	<a href="#">BP2468</a>	Virulence protein, vrg-6
<i>B. pertussis</i>	CDS	<a href="#">BP2627</a>	autotransporter (pseudogene)
<i>B. pertussis</i>	CDS	<a href="#">BP2667</a>	adhesin, fhaS
<i>B. pertussis</i>	CDS	<a href="#">BP2738</a>	autotransporter (pseudogene), bapC

Out of interest, have a look at the Pfam/product browsable catalogues. The Pfam domain of interest is 'Autotransporter beta-domain (PF03797)'. Can you identify autotransporters across the 3 genomes that way?

Note that you can also download this list inc. the sequences by clicking on the 'Report Download' button.

**GeneDB** GeneDB Id List Information 

---

Go To Organisms Go To Shortcuts [Help](#)

---

**ID List**

Please enter your gene/fna ids either:

- as database cross-references (eg in the form GeneDB\_Spombe:SPAC1002.09c).
- or you can just use the systematic ids eg (Tb927.1.710) but in this case you must also set the default organism.

Default Organism: B. pertussis

GeneDB\_Bpertussis:BP0216  
 GeneDB\_Bpertussis:BP0529  
 GeneDB\_Bpertussis:BP0758  
 GeneDB\_Bpertussis:BP0760  
 GeneDB\_Bpertussis:BP0761  
 GeneDB\_Bpertussis:BP0762  
 GeneDB\_Bpertussis:BP0763  
 GeneDB\_Bpertussis:BP0874  
 GeneDB\_Bpertussis:BP1054  
 GeneDB\_Bpertussis:BP1110  
 GeneDB\_Bpertussis:BP1112  
 GeneDB\_Bpertussis:BP1200  
 GeneDB\_Bpertussis:BP1201  
 GeneDB\_Bpertussis:BP1251  
 GeneDB\_Bpertussis:BP1344  
 GeneDB\_Bpertussis:BP1610  
 GeneDB\_Bpertussis:BP1660  
 GeneDB\_Bpertussis:BP1767  
 GeneDB\_Bpertussis:BP1793  
 GeneDB\_Bpertussis:BP1884

---

**Information Required For Each RNA/CDS**

☒ Descriptions

☐ DNA (Unspliced sequence) or cDNA

☐ DNA (Spliced sequence)

☐ Protein sequence

☐ Intergenic Sequence (3')

☐ Intergenic Sequence (5')

☐ Sequence Range

☐ Orthologues

Number of bases: 20

3' distance: 0 5' distance: 0

Submit Query Reset

**Table 2** Autotransporters encoded in the genomes of *B. pertussis*, *B. parapertussis* and *B. bronchiseptica*

	<i>B. pertussis</i>	<i>B. parapertussis</i>	<i>B. bronchiseptica</i>
SphB1	BP0216	BPP0417	BB0419
Novel	-	BPP0449	BB0450
Novel	BP0529	BPP0452	BB0452
Novel	-	BPP0735	BB0821
Novel	-	BPP0822	BB0916
Serum-resistance protein	BP3494	<u>BPP0867</u>	<u>BB0961</u>
Pertactin	BP1054	BPP1150	BB1366
BapA/AidB	BP2224	<u>BPP2251</u>	BB1649
Vag8	BP2315	BPP2415	BB1864
BapC	BP2738	BPP2591	BB2033
Phg	BP1767	BPP1998	BB2246
Novel	<u>BP1793</u>	BPP2022	BB2270
SphB3 (serine protease)	BP1110	BPP2053	BB2301
Novel	<u>BP2627</u>	BPP1256	BB2324
SphB2	BP1660	<u>BPP2745</u>	BB2741
Novel	BP1344	<u>BPP2678</u>	BB2830
Novel	<u>BP1610</u>	BPP2975	BB2941
Novel	-	BPP1618	BB3110
Novel	-	BPP1617	BB3111
TcfA	BP1201	-	BB3291
BapB	BP1200	BPP1815	BB3292

Pseudogenes are underlined.

This is what your completed list should look like.

**Exercise 5.5** Now imagine you're particularly interested in the genomic loci around BB0916 as it only appears to have an orthologue in *B. parapertussis* and not *B. pertussis*. As briefly mentioned earlier, GeneDB supports an Artemis applet with which you can view sequence in more detail as well as being able to download sequence.

***Bordetella bronchiseptica* GeneDB**

This page provides access to the annotation and sequence of *B. bronchiseptica* strain RB50. This is the result of a collaboration of the Sanger Institute with [Duncan Maskell](#) and [Andrew Preston](#) of the Centre for Veterinary Science, [Dept. of Clinical Veterinary medicine](#), The University of Cambridge.

This data described in: Parkhill *et al* (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. [Nature Genetics](#) DOI 10.1038/Ng1227. ([PDF version of this article](#))

Database Entry Point Go To Organisms

Search for **gene**  
by ID/description  
0916  
☒ Include description  
☒ Add wildcards  
  
Full Content Search

**CDS: BB0916**

Search for Go To Organisms Go To Shortcuts [Help](#)  
[Contact curator](#)

**General Information** [Add to Basket](#) [View Basket](#)

**Name**  
**Systematic Name** BB0916  
**Product** autotransporter ([12 others](#))  
**Type** CDS  
**Sequence** [DNA](#) and [Protein](#)

**Location**  
**Chromosome** 1  
**Contig** BB  
**Location** complement(984031..988248) Length: 4218 bp  
**Exons** complement(984031..988248) (Spliced length: 4218 bp)

[Graphical Display \(in Artemis\)](#)

**Context Map:**

BB0910 BB0911 BB0912 BB0913 BB0914 BB0915 >BB0916< BB0917 BB0918 BB0919 BB0920 BB0921 BB0922 BB0923 BB0924 BB0925

Primary Annotation

Hosted by the [Sanger Institute](#)



## Identifying monosaccharide transporters within the genomes of *P. falciparum*, *P. Berghei* and *P. chabaudi*

1

The following exercises aim to introduce you to the features that allow quick and convenient data mining from GeneDB, and will equip you with the tools to use the database to facilitate your own research. Hopefully, they will also make you aware of its strengths and limitations, and highlight the advantage of using several search strategies.

The aim is to identify monosaccharide transporters in *Plasmodium falciparum*, *Plasmodium berghei* and *Plasmodium chabaudi*. Glucose transporters are promising drug targets as asexual-stage parasites depend heavily upon glucose for energy (Joet *et al.*). Compounds such as O-methyl derivatives of glucose, have been shown to selectively inhibit glucose transport by *Plasmodium falciparum*. Hexose Transporter (PfHT). Using a variety of tools/methods, some of which you will already have covered in earlier modules, you'll identify this gene in *Plasmodium falciparum* and then go on to identify its putative orthologues in *P. berghei* and *P. chabaudi*. This would obviously be of interest to a researcher in the field who wanted to assess how similar the putative homologues were to the gene in *Plasmodium falciparum*.

The following are key references that will be provided to you to give you some background information. They are only for your reference and for the purpose of this exercise reading the abstract is probably sufficient.

Joet *et al.*, Comparative characterisation of hexose transporters of *Plasmodium knowlesi*, *Plasmodium yoelli* and *Toxoplasma gondii* highlights functional differences within the apicomplexan family. PMID12238947. Biochem J. 2002 (Dec) pp 923-9.

Krishna *et al.*, Transport processes in *P. falciparum*-infected erythrocytes: potential as new drug targets. PMID:12435441. Int J Parasitol (Dec) pp 1567-73.

### Exercise 1 Searching GeneDB using simple keyword searches

2 Go to the GeneDB homepage (<http://www.genedb.org>)

3

Select *P. falciparum* from the menu

GeneDB

Welcome to the GeneDB website  
Version 2.1

The Wellcome Trust  
Sanger Institute  
Pathogen Sequencing Unit

Database Entry Point

Searches

Search for gene by description in organisms

☒ Include description in search  
☐ Add wildcards to search term

Search Reset

Sequence Searches

[omniBLAST](#)

(Multi-organism BLAST)

Go To [single organism BLAST](#): Choose...

Datasets

Fungi

Go To Choose...

Protozoa

Go To P. falciparum

Parasites

Go To P. falciparum

Viruses

Go To Choose...

Go to our [main search page](#), [complex querying page](#), [AmiGO](#) or our [ID List Download](#)

Information

[Guide to GeneDB](#)

Links

PSU Sequencing Projects

[Prokaryotes](#)

**Plasmodium falciparum GeneDB**

Curated annotation of the *Plasmodium falciparum* 3D7 genome is available from this database.

**Database Entry Point** Go To

Search for  by ID/description  
  
☒ Include description  
☒ Add wildcards

Full Content Search

**Searches/Analysis**  
[omniBLAST](#)  
[BLAST](#)  
[Motif Search](#)  
[EMOWSE](#)  
[AmiGO](#)  
[List Download](#)  
[Cross-Organism Search Page](#)  
[Complex/Boolean Query](#)

**Browse Catalogues**  
[Products](#)  
[Plan](#)  
[InterPro](#)  
[Genome Browser](#)  
[Contig/Chromosome Maps](#)

**Information**

updated 2 May 2005  
[Modified gene models since publication \(Oct 2002\)](#)  
[New gene models since publication \(Oct 2002\)](#)  
[Gene models removed since publication \(Oct 2002\)](#)

[Genome Stats Overview](#)  
[WTSI Plasmodium falciparum project page](#)  
[FTP download \(Sanger\)](#)  
[FTP download \(TIGR\)](#)

Feedback: [Curator](#), [Technical](#)

**Links**

The Plasmodium genome resource [PlasmoDB](#)  
 Functional Genomics [Welcome Trust Functional Genomics Initiative](#)

**News**

2nd May 2005 ([version 2](#))  
 Genome wide review of gene prediction using:

- SNAP gene prediction algorithm
- conserved synteny with *P. knowlesi* and *P. yoelii*
- All splice sites checked
- [tab delimited](#) and [excel annotation summary](#)
- [FTP site updated](#)

5 The results for a search using ‘**pfHT**’ should take you to the gene page for **PFB0210c (next page)**, while the other searches if there are multiple results will Display a results list.

-132-

- 6 Click on the links on the feature page to see how the data are cross-linked and referenced.

**Navigation bar pull down menus:** You can navigate between different organism datasets and search tools using pull down menus

**Gene name and product information:** The description lines are standardized and indexed so that features sharing the same description lines can be retrieved. Access to the nucleotide and amino acid sequences of the feature are also provided.

**Basic location information and context map:** Clicking on the 'Graphical display in Artemis' open up an Artemis applet – which will be discussed further. The applet allows the feature to be viewed in the context of the sequence and additional annotation, such as UTRs

**Details of protein domains defined by Pfam, Interpro, PRINTS, SMART, PROSITE, TIGRFAM, with links to annotation of these families.**

**Gene Ontology associations:** Links will take you to the descriptions of the terms as well as other proteins annotated to the same ontology node.

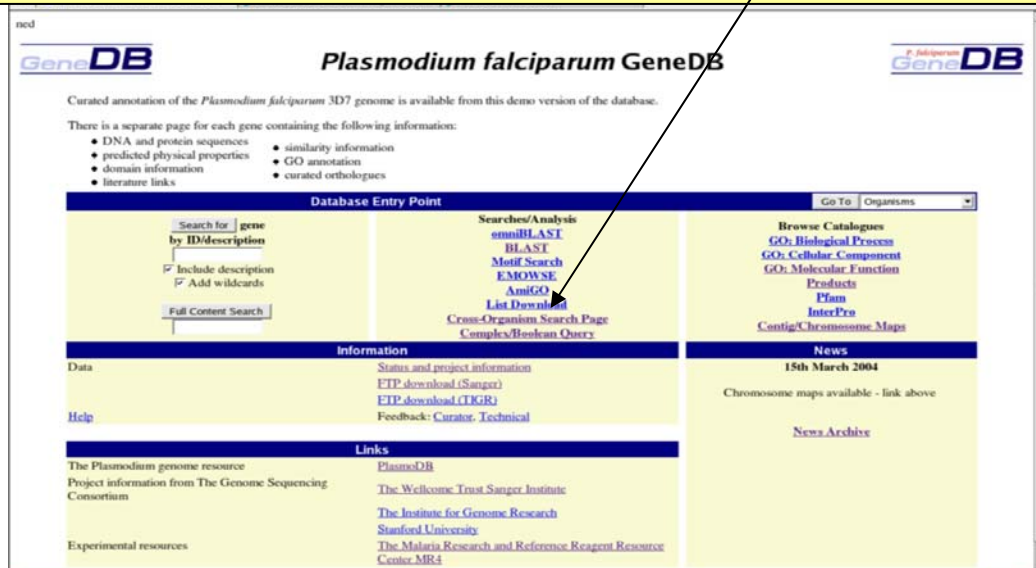
- 7 **Reminder:** Return to box 4 to try the other simple keyword searches if you haven't already. Does PFB0210c appear in the results from these searches. Also try entering the search terms in the **Full Content Search** box to assess how this affects the results. Use quotes to group two search terms into a phrase, e.g. "hexose transporter" looks for the occurrence of these two terms together.

Through this exercise you should have seen that when searching by geneID/description the gene is found when using, **PfHT**, monosaccharide transporter but not when using sugar transporter, glucose transporter or fructose transporter. **PfHT** is a gene name and because this has been annotated into the database from the literature it is detected in the database. The other search terms are descriptions of the product of the gene and although the protein can transport fructose and hexose it is described as a monosaccharide transporter in the database. This is just an example of how the way in which a gene is described can affect the results of simple searches. It is always better to try several search terms and compare the results. Doing a full content search will search all of the annotation fields associated with a gene. Thus it tends to be comprehensive but give many hits.

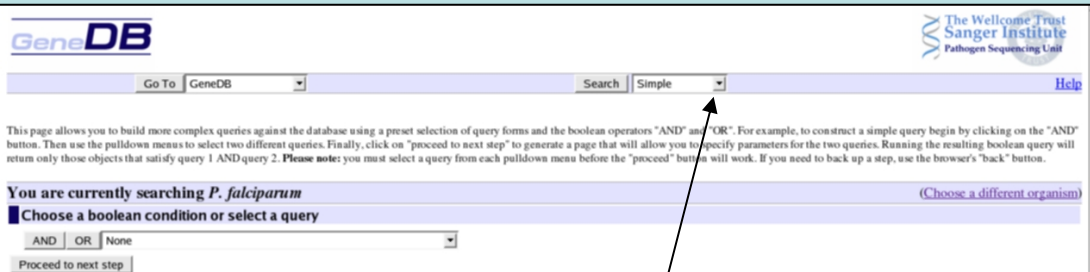
- 8 We're now going to move onto complex querying which allows searching of several genomes concurrently if desired, and allows a diverse range of queries to be used.

## Exercise 2 Searching of multiple genomes using more complex keywords, manipulation of and downloading results sets.

- 9 Complex querying of multiple genomes. Click on the complex querying link.

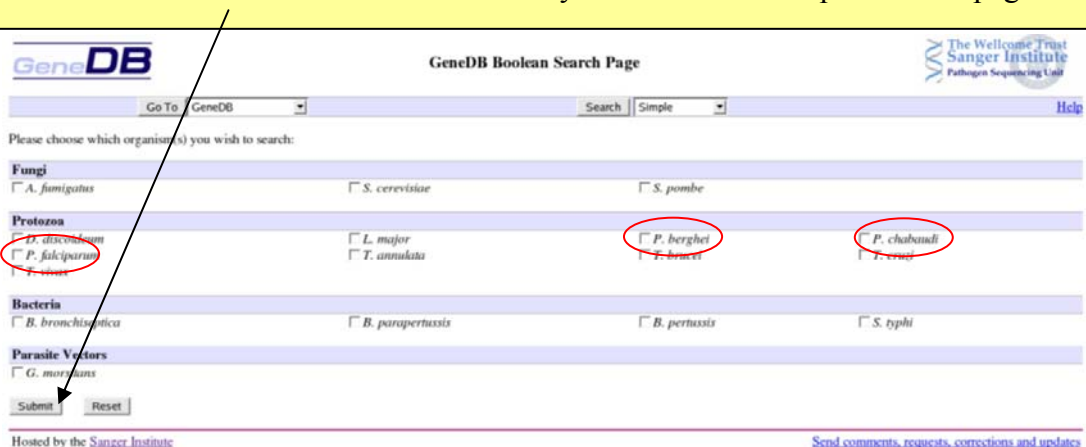


- 10 This will bring you through to the page below. Since you are going to apply the search to all three Plasmodium genomes in GeneDB, you have to select a **complex** search.



- 11 Select *complex* from the search drop down menu which will take you through to the following page.

- 12 Click on the boxes for *P. falciparum*, *P. berghei* and *P. chabaudi* (circled red). Then click the submit button. This will take you back to the complex search page.



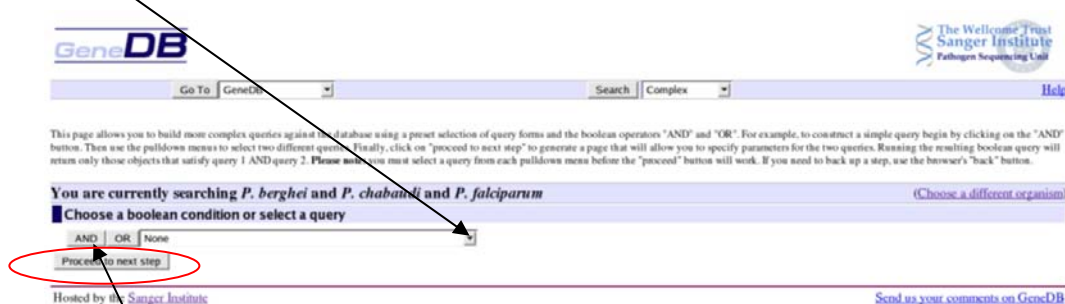


- 13 Complex queries can be built up using this page. It uses a Boolean approach. Many different data types can be search and **AND** or **OR** can be used to enhance searching.

- 14 Our objective is to find hexose transporters in the three *Plasmodium* genomes. Try the following complex search strategies and compare the results you obtain. The following pages show you how to set up the first query. Then try numbers 2 and 3 for yourself.

- 1 (Proteins with a product containing a particular keyword or phrase: *transporter*) **AND** (GO component: *membrane*).
- 2.(Proteins containing one or more transmembrane domains, try between 8 and 14) **AND** (Proteins with a particular GO process: *transport*).
- 3.(Proteins with a product containing a particular keyword or phrase: *transporter*) **AND** (Proteins with a signal peptide).

- 15 Click on the AND button and a second pull down menu should appear below the first. Select *Proteins with a product containing a particular keyword or phrase* by clicking on the first pull-down menu and selecting this option.



- 16 In the second pull down menu, select the option *Proteins with a predicted GO component*. Then click on the proceed to next step button (circled red). The screen should appear as below.





17

Type transport in the Keyword box (underlined red)

Then select use the scroll bar to scroll down the GO term until you get to membrane and then select it. Then click the submit query button (circled red).

You are currently searching *P. falciparum* (Choose a different organism)

AND

Proteins with annotation matching a particular keyphrase

Keyword: transport ?

Proteins with a predicted GO component

GO Component: 26S proteasome, 6-phosphofructokinase complex, acetyl-CoA carboxylase complex, actin capping protein complex, actin cytoskeleton, actin filament, alpha DNA polymerase:primase complex, alpha-ketoglutarate dehydrogenase complex (sensu Eukarya), anaphase-promoting complex, apical complex ?

Query options

Rows per page: 20 ?

Run query

Submit (circled red) Reset

Hosted by the Sanger Institute Send us your comments on GeneDB

18

Select rows per page to 100 from the drop down menu.

Then select use the scroll bar to scroll down the GO term until you get to membrane and then select it. Then click the submit query button (circled red).

19

The query should return 74 results. Examine the results to see which proteins could be hexose/glucose/monosaccharide transporters and whether PFB0210c is present. Then use the browser back button to go back and submit the second and third queries. Note that the results page tells you whether the protein has been manually or automatically annotated.

20

We're now going to look at ways of treating the results sets that are obtained from our Boolean searches. We can add, subtract and intersect different results sets using the history page. We'll also look at downloading the results sets.

GeneDB Results 1 - 13 of 13

Go To GeneDB Search Simple Hg

Query: (Genes for *pberghet:pchabaudi:malaria*) intersect ((Proteins with a product containing the keyword or phrase 'transporter') intersect (Proteins containing a predicted signal peptide.))

[1-13]	name	id	organism	description	[1-1]
1		PB000033.01.0	<i>P. berghet</i>	transporter protein, putative	Automatic Annotation
2		PC000560.00.0	<i>P. chabaudi</i>	phosphate transporter, putative	Automatic Annotation
3		PC000437.02.0	<i>P. chabaudi</i>	transporter protein, putative	Automatic Annotation
4		PB000168.03.0	<i>P. berghet</i>	phosphate transporter, putative	Automatic Annotation
5		PC000604.03.0	<i>P. chabaudi</i>	sugar transporter, putative	Automatic Annotation
6		PF11_0141	<i>P. falciparum</i>	UDP-galactose transporter, putative	Manual Annotation
7		PEF0450c	<i>P. falciparum</i>	putative transporter protein	Manual Annotation
8		PE07_0065	<i>P. falciparum</i>	zinc transporter, putative	Manual Annotation
9		PE07_0070	<i>P. falciparum</i>	transporter/permease protein, putative	Manual Annotation
10		MAL13P1.206	<i>P. falciparum</i>	phosphate transporter, putative	Manual Annotation
11		PE14_0133	<i>P. falciparum</i>	ATP-dependent transporter, putative	Manual Annotation
12		PEC0125w	<i>P. falciparum</i>	ABC transporter, putative	Manual Annotation
13		PEC0875w	<i>P. falciparum</i>	transporter, putative	Manual Annotation

To view this result set later, or to download this data [visit the history page](#).

[Start a completely new complex query](#)

21

Click on the History page link to bring up details of search results.

22

The History page allows results sets to be viewed and downloaded. It is only active for Boolean searches, but is a very useful way of tracking and manipulating results sets. It also allows results sets to be added together (union), the contents of one set removed from another (subtract) and identification of those that appear in both sets (intersect)

GeneDB Query History

Go To: GeneDB Search: Simple Help

This page displays results of queries executed using the [boolean](#) query pages. The page will remain empty until searches have been run. Please note that a search combining boolean operators will return multiple results files. Chosen queries are first executed across all organisms within GeneDB and subsequently narrowed down to only return data from the organisms of choice. Once query results have been retrieved, users can combine results files in one of three ways: adding files together (union), identifying common results between files (intersect), or identifying unique results between files (subtraction). **Please note:** you must have cookies enabled in your browser in order for this page to work correctly.

The result files can either be viewed or downloaded as a FASTA file of DNA or protein sequences.

Query	Start time	Response time	Result Download Size
<input type="checkbox"/> (Proteins with a product containing the keyword or phrase 'transporter') <b>intersect</b> (Proteins predicted to have GO component 'membrane')	5:06:08 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 438
<input checked="" type="checkbox"/> (Genes for <i>pberghei:pchabaudi:malaria</i> ) <b>intersect</b> ((Proteins with a product containing the keyword or phrase 'transporter') <b>intersect</b> (Proteins predicted to have GO component 'membrane'))	5:06:08 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 67
<input type="checkbox"/> (Proteins predicted to have between 8 and 14 transmembrane domains.) <b>intersect</b> (Proteins predicted to have GO process 'transport')	5:10:43 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 428
<input checked="" type="checkbox"/> (Genes for <i>pberghei:pchabaudi:malaria</i> ) <b>intersect</b> ((Proteins predicted to have between 8 and 14 transmembrane domains.) <b>intersect</b> (Proteins predicted to have GO process 'transport'))	5:10:42 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 42
<input type="checkbox"/> (Proteins with a product containing the keyword or phrase 'transporter') <b>intersect</b> (Proteins containing a predicted signal peptide.)	5:12:06 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 303
<input type="checkbox"/> (Genes for <i>pberghei:pchabaudi:malaria</i> ) <b>intersect</b> ((Proteins with a product containing the keyword or phrase 'transporter') <b>intersect</b> (Proteins containing a predicted signal peptide.))	5:12:06 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 13

Follow one of the links in the table above to view a query result set or select two of the result sets and use one of the following buttons to combine them. "Union" will create a result set that contains all the genes in either of the selected sets. "Intersect" will create a result set that contains only the genes in both of the selected sets. "Subtract" will remove any genes in the second (i.e., appearing lower in the list) set from the first.

[UNION](#) or [INTERSECT](#) or [SUBTRACT](#) the selected query results (a new entry will appear at the end of the list.)

23

Click on the boxes for the results sets for query 1 and query 2 (marked by the red arrows) and then click on intersect (circled red). A new results set will appear containing only the genes which occur in the results sets for query 1 and 2. **Why doesn't the gene for the putative monosaccharide transporter in *Plasmodium chabaudi* (PC000736.00.0) appear in the results set for query 2?**

24

We're now going to look at how to download a results set and see what formats and different parts of the dataset can be obtained by choosing different options.

GeneDB Query History

Go To: GeneDB Search: List Download Help

This page displays results of queries executed using the [boolean](#) query pages. The page will remain empty until searches have been run. Please note that a search combining boolean operators will return multiple results files. Chosen queries are first executed across all organisms within GeneDB and subsequently narrowed down to only return data from the organisms of choice. Once query results have been retrieved, users can combine results files in one of three ways: adding files together (union), identifying common results between files (intersect), or identifying unique results between files (subtraction). **Please note:** you must have cookies enabled in your browser in order for this page to work correctly.

The result files can either be viewed or downloaded as a FASTA file of DNA or protein sequences.

Query	Start time	Response time	Result Download Size
<input type="checkbox"/> (Proteins with a product containing the keyword or phrase 'transporter') <b>intersect</b> (Proteins predicted to have GO component 'membrane')	5:06:08 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 438
<input checked="" type="checkbox"/> (Genes for <i>pberghei:pchabaudi:malaria</i> ) <b>intersect</b> ((Proteins with a product containing the keyword or phrase 'transporter') <b>intersect</b> (Proteins predicted to have GO component 'membrane'))	5:06:08 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 67
<input type="checkbox"/> (Proteins predicted to have between 8 and 14 transmembrane domains.) <b>intersect</b> (Proteins predicted to have GO process 'transport')	5:10:43 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 428
<input type="checkbox"/> (Genes for <i>pberghei:pchabaudi:malaria</i> ) <b>intersect</b> ((Proteins predicted to have between 8 and 14 transmembrane domains.) <b>intersect</b> (Proteins predicted to have GO process 'transport'))	5:10:42 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 42
<input type="checkbox"/> (Proteins with a product containing the keyword or phrase 'transporter') <b>intersect</b> (Proteins containing a predicted signal peptide.)	5:12:06 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 303
<input type="checkbox"/> (Genes for <i>pberghei:pchabaudi:malaria</i> ) <b>intersect</b> ((Proteins with a product containing the keyword or phrase 'transporter') <b>intersect</b> (Proteins containing a predicted signal peptide.))	5:12:06 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 13
<input type="checkbox"/> ((Genes for <i>pberghei:pchabaudi:malaria</i> ) <b>intersect</b> ((Proteins with a product containing the keyword or phrase 'transporter') <b>intersect</b> (Proteins predicted to have GO component 'membrane')))) <b>intersect</b> ((Genes for <i>pberghei:pchabaudi:malaria</i> ) <b>intersect</b> ((Proteins predicted to have between 8 and 14 transmembrane domains.) <b>intersect</b> (Proteins predicted to have GO process 'transport'))))	5:29:52 PM	<1 second	<a href="#">view</a> <a href="#">Download</a> 19

Follow one of the links in the table above to view a query result set or select two of the result sets and use one of the following buttons to combine them. "Union" will create a result set that contains all the genes in either of the selected sets. "Intersect" will create a result set that contains only the genes in both of the selected sets. "Subtract" will remove any genes in the second (i.e., appearing lower in the list) set from the first.

[UNION](#) or [INTERSECT](#) or [SUBTRACT](#) the selected query results (a new entry will appear at the end of the list.)

25

Click on the download link for the new results set that you generated by the intersection of queries 1 and 2.

26

In the initial download window the description of the genes in the results list will appear. The lower part of the window boxed is used to select what type of information you want to download. These options include the DNA sequence with introns (unspliced) or without introns (spliced), the protein sequence and either 5' or 3' regions flanking the gene to a chosen number of bases. This would be very useful when examining regulatory elements such as promoter regions or UTR (untranslated region)

GeneDB

GeneDB Id List Information

The Wellcome Trust  
Sanger Institute  
Pathogen Sequencing Unit

Go To Organisms Go To Shortcuts Help

Feb. 11th 2004  
Please note the ID list download and the shopping basket are brand new. They were introduced early because of a hardware failure and haven't had the usual levels of testing, so apologies for any usability rough edges or error messages. The data that comes out has been tested and is accurate.

**ID List**

Please enter your gene/ma ids either:

- as database cross-references (eg in the form GeneDB\_Spombe:SPAC1002.09c).
- or you can just use the systematic ids eg (Tb927.1.710) but in this case you must also set the default organism.

Default Organism: None

GeneDB\_Ffalciparum:FF80210c  
GeneDB\_Fchabaudi:PC000722.01.0  
GeneDB\_Ffalciparum:FF07\_0070  
GeneDB\_Ffalciparum:FF14\_0679  
GeneDB\_Fchabaudi:PC000560.00.0  
GeneDB\_Pberghei:PB000562.01.0  
GeneDB\_Pberghei:PB000168.03.0  
GeneDB\_Ffalciparum:FF11\_0141  
GeneDB\_Fchabaudi:PC000144.02.0  
GeneDB\_Fchabaudi:PC000134.02.0  
GeneDB\_Pberghei:PB000598.03.0  
GeneDB\_Pberghei:PB000201.02.0  
GeneDB\_Ffalciparum:FFE0825w  
GeneDB\_Ffalciparum:FFE1185w  
GeneDB\_Ffalciparum:FFE1455w  
GeneDB\_Ffalciparum:FFE0450c  
GeneDB\_Ffalciparum:MAL8P1.32  
GeneDB\_Pberghei:PB000569.02.0  
GeneDB\_Pberghei:PB000033.01.0

**Information Required For Each RNA/CDS**

☒ Descriptions  
☐ DNA (Unspliced sequence)  
☐ DNA (Spliced sequence)  
☐ Protein sequence  
☐ Intergenic Sequence (3')  
☐ Intergenic Sequence (5')  
☐ Sequence Range

Number of bases: 20  
3' distance: 20 5' distance: 20

Submit Query Reset

Hosted by the Sanger Institute [Send comments, requests, corrections and updates](#)

27

For the sake of our exercise lets assume that we will proceed to compare the amino acid sequences of our transporters. Thus we would click on Protein sequence (arrow) and then on the submit query button (circled). Then to download this information you would save the page from your browser. This is not described here as the process will be particular to the internet browser that your using.

### Exercise 3 Search strategies using omniBLAST and browsing of the Pfam domain catalogue

28 We are now going to return to consider and run a few other search strategies which make use of the strengths of GeneDB.

1. Use of a text keyword search across several organisms using the cross-organism search page. This can be a quick and powerful way to identify genes/proteins in other organisms that perform very similar functions to your gene of interest. This can be achieved with simple keywords and requires little previous knowledge about the gene of interest. Once a gene or protein has been found that meets the keyword criteria, e.g. sugar+transporter, the DNA or Protein sequence can be searched against any genome in GeneDB using omniBLAST. In our case we would want to search the three *Plasmodium* genomes.
2. Another powerful approach makes use of the fact that many protein domains that are diagnostic of a particular function have already been characterised and assigned to many genes within the database. Thus, if we know that our gene of interest has a particular Pfam or Interpro domain then we can browse through the Pfam or Interpro catalogue for genes which have this domain. This can be done concurrently for several organisms using the Cross-Organism search page.

29 Select the Cross-Organism Search Page. The link can be found on the *Plasmodium falciparum* main page or any of the organism main pages. See the figure under **box 4**. In the full text search section (underlined red) enter "*hexose transporter*" + "*glucose transporter*" (include the quotes). In the adjacent box circled red select **All organisms** from the pull down menu. Then click on the adjacent Search button. (red arrow)

The screenshot shows the GeneDB Search Page with the following sections:

- GeneDB Search Page** (Header)
- Searching By Name/Id/Description** (Section 1)
- Full-text search (site-wide)** (Section 2)
- Browsing By Products/Description** (Section 3)
- Browse By SWISS-PROT Keywords** (Section 4)
- Pfam Assignments** (Section 5)
- InterPro Assignments** (Section 6)
- Browsing By Ryley Catalogue** (Section 7)

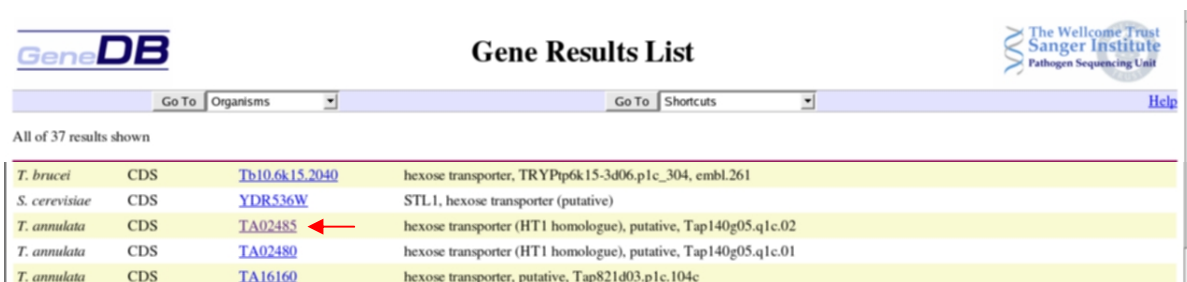
In the **Full-text search (site-wide)** section, the search input field contains the text: "hexose transporter" + "glucose transporter". To the right of the input field is a dropdown menu currently set to "P. falciparum", which is circled in red. A red arrow points to the "Search" button next to the dropdown.

30

A results list will then be displayed as below (note that the whole list is not shown). We are now going to take the protein sequences of one of these genes and search for similar protein sequences in the genomes of *P. falciparum*, *P. berghei* and *P. chabaudi* using Omnitblast. This uses the wublastp algorithm. The example below will take you through this process for one protein, but if you have time try one or two others. To start with use the sequence of a protein from *T. annulata* (TA02485) as this has been annotated as a hexose transporter homologue (HT1 homologue) which is the name of the gene in *P. falciparum* as it appears in the literature, and so it is likely that it is an orthologue of the protein in *P. falciparum*.

31

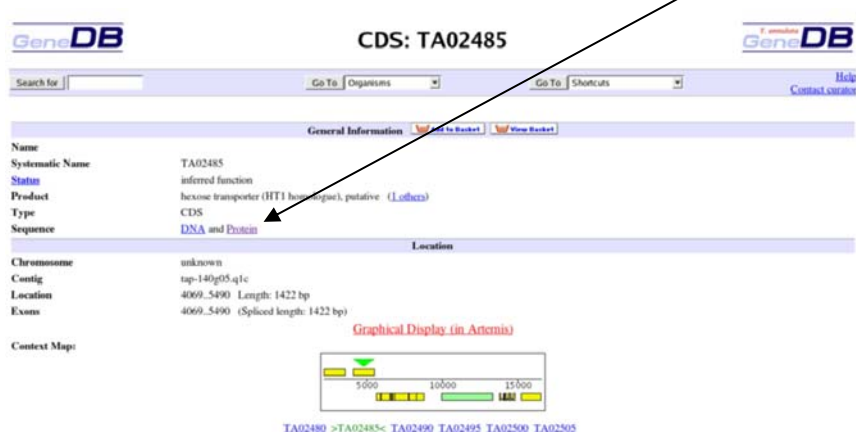
Click on the link for as shown below. This will take you through to the feature page for this gene.



Organism	Feature Type	Feature ID	Description
<i>T. brucei</i>	CDS	<a href="#">Tb10.6k15.2040</a>	hexose transporter, TRYPIP6k15-3d06.p1c_304, embi.261
<i>S. cerevisiae</i>	CDS	<a href="#">YDR536W</a>	STL1, hexose transporter (putative)
<i>T. annulata</i>	CDS	<a href="#">TA02485</a>	hexose transporter (HT1 homologue), putative, Tap140g05.q1c.02
<i>T. annulata</i>	CDS	<a href="#">TA02480</a>	hexose transporter (HT1 homologue), putative, Tap140g05.q1c.01
<i>T. annulata</i>	CDS	<a href="#">TA16160</a>	hexose transporter, putative, Tap821d03.p1c.104c

32

To access the protein sequence to do an Omnitblast search click on the protein link.



**GeneDB CDS: TA02485**

Search for:  Go To: Organisms Shortcuts Help Contact curator

**General Information** [Add to Basket](#) [View Basket](#)

Name: TA02485  
 Systematic Name: TA02485  
 Status: inferred function  
 Product: hexose transporter (HT1 homologue), putative (1.08kcs)  
 Type: CDS  
 Sequence: [DNA](#) and [Protein](#)

**Location**

Chromosome: unknown  
 Contig: tap-140g05.q1c  
 Location: 4069..5490 Length: 1422 bp  
 Exons: 4069..5490 (Spliced length: 1422 bp)

Context Map:

Graphical Display (in Artemis)

TA02480 >TA02485< TA02490 TA02495 TA02500 TA02505



33

Click on the Send to GeneDB omniBLAST link to send this sequence as a query to the GeneDB omniBLAST server.

GeneDB Sequence

Gene: TA02485

Search for: [ ] Go To: Organisms [ ] Go To: Shortcuts [ ] Help Contact Curator

Go to main detail page for TA02485

Unspliced DNA

Send to GeneDB omniBLAST Send to GeneDB BLAST Send to BLAST at NCBI

>TA02485 |||hexose transporter (HT1 homologue), putative|Theileria annulata|chr unknown|||Manual

ATGAAGCTTA AGCATCATCAT ACCTTTGATT GCAGGGGGA GCATTGAGC TCTTGAGCT

TAATCTTTTC GCTTACTCT GCAGGCTTC AATACATCA AGCAATTCG TATCTAGAT

ATGGAATGCT CTAAGGAGA GACATATATT TTGATTCG CCAATCTCA ATCTTTGCT

GCTCTTCCA ATGCTTTAC TTCTTTGGA GCAGCTTTC GTTCACTTC GATTGAGCT

TGCGAAGAA TGGAGGAGC AGCTACATTC ATGCTCTTA ATGCTTTT GTTTGAGA

TCTGATATAT CAGCTTCATC TCTTCATTTT GCTATCTGT TCTAGAGCTG TTTGATTT

GCTTTGGA TTGTTTGGC GCTCTCAAT CAGCTTTT TGTGTAGAT ATGTCACT

AGCAAGAA AATACCTTC ACTATATAT CAGTTTCA TACATCTTC TTACTCAT

TGACAGGAT GCAATAGC CCACTCTCT CTACACAA ACCTGCTAGC GAGCAAGC

TTTAACCTTA CTACATGGA TAGATTCTA TGGTACTA CACATCTCT ACAGTTCTA

TCTGTCTAT CTGATGAT TCTATCTTC CATCTTAC ACATATGAA

TTGATTTCAA AGGCAAAAC AGAGAGCT CCAACCTTA TAGAAGCT TCATGGAAG

GAGAGCTTC ATCAAGTATA TGATGAATTC GAGAGATC AGAGAGCT TCATCAAC

CGAGCATAC CACTATATAT AGCTTTGGA ATGCTCAAT AGAGAGCT GATATACAT

CGATTCTTT TGGCTCAGT ACACCAATTC GAGCTCTAA CCACTCTAC CTCAAATTC

ACAAATTT TCTTAAGT AATGCTGGA ATTCAACT CACCTTAGC TTCACTCT

ATCTTTTTC TAACTCTCT TGGCTCACT TCACTTAGC TAACTTTG

AGGAACCT TCTGCTTC GGTATAGG TTTCTACAT TTTTATGCT CCACTCT

TTTCAACG CATTGGAAG AGAGAGCT TGGCTCTCT TCTATCTC ATAGCTTC

TTGCTTTA TATGATAT TGATTTGGA TATGCTGCT TATTTATG

GAGCTATTG CACTGAATA CAGGATGGA GCTTAAAGT TGGCTCTT CATTAATG

CTATCTGCT GACTCACTT GATTGCTCA GAATTTTGA TTTCTACT TCAGCTCT

CTAGCTACT TCTCTCTG CTCTCTCT TCTGCTTA TTTATGAT ACTCTCAT

AAGAACAA AGGCTTCT GATCGAAA CATATGAT AA

Protein

Send to GeneDB omniBLAST Send to GeneDB BLAST Send to BLAST at NCBI

>TA02485 |||hexose transporter (HT1 homologue), putative|Theileria annulata|chr unknown|||Manual

MRKASPLI AGASGALG LQCLPSAL NYDEPATO MEKQKETT TCKPSLGA

GLNCTFLG AAFCSLLGL SKIGRRVTL WYNWFFVC SVLSASSVF AMPLGLKLS

GFGLAAVI PVFLVEIHP KRRYFATY QLTFTGLLI SAGWLAHR VYKLVSGE

RLTYDRPV RSTQPLPV CNAALILDE PVVFPDPV LSGKTEEA RYVTRJAG

REYREYDF DRQVAKST PSLILLALG NKYRKYIDH APVLAQQL VQVTLTSNV

TKLFLAYNR WYNLTASS ILNYPVAT ILTYLQKFS RTTLYVGLG PFTFMALPS

YAFPLGAS WYVYSISQ PVFLFAPG LSGDMPTG IVTATYRDS ALSNVTIN

LOSLTLIAS RFLIYISVIL VFTLFAPL FGTIVIVFT KETKVAIGK AYD

Go to main detail page for TA02485

Hosted by the Sanger Institute Curator feedback Technical feedback

34

The sequence will automatically be put into the query box (red underlined) and the protein option selected (red circled). Choose the genomes that you want the query to be searched against (note that only the top part of the page is shown). To go down to protozoan genomes click on the jump down to Protozoa link (red arrow). Note that the *T. annulata* sequence will automatically be selected. Select the protein databases for *P. falciparum*, *P. chabaudi* and *P. berghei* by clicking on the small text box adjacent to **P**, then click on the start omniBLAST button (red box).

GeneDB omniBLAST Server

Go To: Organisms [ ] Go To: Shortcuts [ ] Help

OmniBLAST will perform a BLAST search on a set of protein databases (BLASTP or BLASTX, depending on the query sequence) or nucleotide databases (BLASTN and TBLASTX or TBLASTN) available in GeneDB and return a list of the best five HSP for each database. If there are any HSP you can click on Full Search to see the complete BLAST output.

To search individual databases with different parameters Go To: single organism BLAST: [Choose...]

QUERY DATA

Paste your sequence here.  
fasta format or just plain text will do

>TA02485 |||hexose transporter (HT1 homologue), putative|Theileria annulata|chr unknown|||Manual

MRKASPLIAGASGALG LQCLPSAL NYDEPATO MEKQKETT TCKPSLGA

GLNCTFLG AAFCSLLGL SKIGRRVTL WYNWFFVC SVLSASSVF AMPLGLKLS

GFGLAAVI PVFLVEIHP KRRYFATY QLTFTGLLI SAGWLAHR VYKLVSGE

RLTYDRPV RSTQPLPV CNAALILDE PVVFPDPV LSGKTEEA RYVTRJAG

REYREYDF DRQVAKST PSLILLALG NKYRKYIDH APVLAQQL VQVTLTSNV

TKLFLAYNR WYNLTASS ILNYPVAT ILTYLQKFS RTTLYVGLG PFTFMALPS

YAFPLGAS WYVYSISQ PVFLFAPG LSGDMPTG IVTATYRDS ALSNVTIN

LOSLTLIAS RFLIYISVIL VFTLFAPL FGTIVIVFT KETKVAIGK AYD

Determine sequence type automatically ☐ or set sequence type to DNA ☐ protein ☒

Start omniBLAST Reset

Note: OmniBLAST searches may take several minutes depending on the number of selected databases. Please check the databases chosen below are correct

DATABASE OPTIONS

Search only the BLAST databases selected below

Jump down page to: Protozoa Bacteria Parasite Vectors

A. fumigatus

☐ A. fumigatus finished BAC sequences N

☐ A. fumigatus BAC predicted genes (coding sequences) N

☐ A. fumigatus whole genome shotgun reads N

☐ A. fumigatus BAC predicted proteins P

☐ A. fumigatus BAC ends N

35

Note that OmniBLAST can be used to search on the basis of DNA sequence also. Sequence can also be pasted into the query box (FASTA or plain text) and searched.

36

Click on the Retrieve button (red arrow). The omniBLAST search may take a while depending on the number size of the search. Once completed the omniBLAST results are presented in a summarised format (as shown below) as BLAST output files are large and detailed. The top five hits in each search are summarised. **Does this search detect PFB0210c (PfHT)? Are these results consistent with what you thought were the orthologues of pHT in *Plasmodium berghei* and *Plasmodium chabaudi* from your previous searches?** Once you have looked at the search results try clicking on the various options in the results page.

**GeneDB OmniBLAST Server Submission**

Go To:  Search:  [Help](#)

Retrieve result for id:  [retrieve](#)

Your BLAST query has been added to the queue of jobs.  
The majority of BLASTs are completed within two minutes.

To retrieve your results, click the **retrieve** button above, or use the following URL:  
<http://www.genedb.org/genedb/2blast/getblast?id=s2bA7905L08QeE96d7270>

Hosted by the [Sanger Institute](#) [Send us your comments on GeneDB](#)

---

**GeneDB Blast Server Results**

Go To:  Search:  [Help](#)

Retrieve result for id:  [retrieve](#)

At peak times your BLAST searches could take longer than normal.  
Please be patient.

BLAST results are kept on our servers for three days following query submission. Results may be retrieved any number of times during this period. After this time queries must be resubmitted if further examination is required.

There were no significant matches to Pfam domains

Summary for: *T. annulata* predicted genes (coding sequences) [wutblastn], for query: TA02485 [\[Full BLAST Search\]](#)

Name	Score	(P/N)	N	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA02485</a>	Score: 2398	(P/N): 4.0e-250	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA02480</a>	Score: 2255	(P/N): 5.7e-235	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA16160</a>	Score: 1548	(P/N): 4.7e-160	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA12370</a>	Score: 90	(P/N): 0.054	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA05960</a>	Score: 83	(P/N): 0.29	N: 1	<a href="#">[Full Sequence]</a>

Summary for: *T. annulata* predicted proteins [wublastp], for query: TA02485 [\[Full BLAST Search\]](#)

Name	Score	(P/N)	N	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA02485</a>	Score: 2398	(P/N): 5.9e-251	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA02480</a>	Score: 2255	(P/N): 8.3e-236	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA16160</a>	Score: 1548	(P/N): 6.9e-161	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA12370</a>	Score: 90	(P/N): 4.4e-05	N: 2	<a href="#">[Full Sequence]</a>
Name: <a href="#">TA05960</a>	Score: 83	(P/N): 0.058	N: 1	<a href="#">[Full Sequence]</a>

Summary for: *P. berghei* proteins [wublastp], for query: TA02485 [\[Full BLAST Search\]](#)

Name	Score	(P/N)	N	<a href="#">[Full Sequence]</a>
Name: <a href="#">PB000562.01.0</a>	Score: 687	(P/N): 4.7e-69	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">PB000161.03.0</a>	Score: 138	(P/N): 1.3e-08	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">PB000390.01.0</a>	Score: 100	(P/N): 0.00087	N: 2	<a href="#">[Full Sequence]</a>
Name: <a href="#">PB001144.02.0</a>	Score: 67	(P/N): 0.088	N: 3	<a href="#">[Full Sequence]</a>
Name: <a href="#">PB105950.00.0</a>	Score: 59	(P/N): 0.18	N: 1	<a href="#">[Full Sequence]</a>

Summary for: *P. falciparum* proteins [wublastp], for query: TA02485 [\[Full BLAST Search\]](#)

Name	Score	(P/N)	N	<a href="#">[Full Sequence]</a>
Name: <a href="#">PFB0210c</a>	Score: 704	(P/N): 2.6e-71	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">PFI0955w</a>	Score: 262	(P/N): 7.7e-22	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">PEL0890c</a>	Score: 89	(P/N): 0.015	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">PE11_0310</a>	Score: 90	(P/N): 0.016	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">PFB0465c</a>	Score: 72	(P/N): 0.044	N: 2	<a href="#">[Full Sequence]</a>

Summary for: *P. chabaudi* proteins [wublastp], for query: TA02485 [\[Full BLAST Search\]](#)

Name	Score	(P/N)	N	<a href="#">[Full Sequence]</a>
Name: <a href="#">PC000736.00.0</a>	Score: 443	(P/N): 3.5e-43	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">PC000604.03.0</a>	Score: 257	(P/N): 6.6e-21	N: 1	<a href="#">[Full Sequence]</a>
Name: <a href="#">PC000442.01.0</a>	Score: 230	(P/N): 6.4e-20	N: 1	<a href="#">[Full Sequence]</a>

Annotations:

- Clicking here will show full BLAST results (points to [\[Full BLAST Search\]](#))
- Clicking on the systematic identifier (systematic id) will show the alignment of this protein against the query. (points to [TA02485](#))
- Clicking here will show the full sequence of this protein (points to [\[Full Sequence\]](#))

37

This approach has demonstrated how an omniBLAST search can identify the gene of interest in your organism when a well-annotated orthologue exists in another organism. So this is a useful alternative strategy to searching on keywords alone, which we have seen can in some cases be misleading. It also shows that the full text search (site wide) is a powerful way of searching the annotation of all the genomes in GeneDB for possible orthologues.

- 38 An alternative way of identifying potential orthologues is the presence of a protein domain that is associated with that function. This approach also makes use of the Cross Organism search Page which allows browsing of Pfam and Interpro assignments across several genomes concurrently. Let us assume that our previous searches had uncovered that the Pfam domain *Sugar (and other) transporter* (PF00083) is a Pfam domain associated with Hexose Transporters. Note that in PF00083, PF stands for Pfam. To view more details about this protein domain goto the gene page (see under box 6 for reference) and click on PF00083 which is in a red box in the figure under box 6. We want to search for proteins in *Plasmodium falciparum*, *Plasmodium chabaudi*, *Plasmodium berghei* that also have this domain.

- 39 Go to the Cross-Organism search page (for reference see figures under Box 9 and Box 2 if needed). In the Pfam Assignments section click the boxes for *P. falciparum*, *P. chabaudi*, *P. berghei* and then Browse button (circled red).

The screenshot shows the GeneDB Search Page with the following sections:

- GeneDB Search Page**: Includes navigation links like 'Go To Organisms', 'Go To Shortcuts', and 'Help'.
- Searching By Name/Id/Description**: A search bar with options to 'Include description', 'Add wildcards', and 'Search Names/IDs'. Below this are checkboxes for various organisms: Fungi (A. fumigatus, S. cerevisiae, S. pombe), Protozoa (D. discoideum, L. major, P. berghei, P. chabaudi, T. cruzi, P. falciparum, T. annulata, T. brucei, T. vivax), Bacteria (B. bronchiseptica, B. paraperthussis, B. pertussis, S. typhi), and Parasite Vectors (G. morsitans).
- Full-text search (sitewide)**: A section for searching across the entire site, including wildcards and quotation marks.
- Browsing By Products/Description**: A section for browsing through descriptions/products for various organisms.
- Browse By SWISS-PROT Keywords**: A section for browsing through SWISS-PROT keywords.
- Pfam Assignments**: A section for browsing through Pfam domains. The 'Browse' button is circled in red. Below this are checkboxes for various organisms: Fungi (S. cerevisiae, S. pombe), Protozoa (D. discoideum, L. major, P. berghei, P. chabaudi, T. cruzi, P. falciparum, T. annulata, T. brucei, T. vivax), Bacteria (B. bronchiseptica, B. paraperthussis, B. pertussis, S. typhi), and Parasite Vectors (G. morsitans).
- InterPro Assignments**: A section for browsing through InterPro assignments.
- Browsing By Riley Catalogue**: A section for browsing through the Riley Catalogue.


- 40 In the Pfam list click on 'S' (red arrow) and then select sugar (and other) transporter from the list.

The screenshot shows the Pfam List page with the following sections:


- Pfam List**: Includes navigation links like 'Go To Organisms', 'Go To Shortcuts', and 'Help'.
- Results 1 to 100 of 816 results shown**: A section indicating the number of results.
- Key:** A warning about automatic prediction.
- Previous**: A section for navigating to previous results.
- 0-9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z**: An alphabetical index of protein domains. A red arrow points to the 'S' tab.
- Next 100**: A section for navigating to the next 100 results.
- Protein Domains**: A list of protein domains, including:
  - 'chromo' (CHR) domain (PF00385) (2)
  - 'Cold-shock' DNA-binding domain (PF00312) (1)
  - 1-deoxy-D-xylulose 5-phosphate reductoisomerase (PF02670) (1)
  - 14-3-3 protein (PF00244) (2)
  - 2-oxoacid dehydrogenases acyltransferase (catalytic domain) (PF00198) (3)
  - 2Fe-2S iron-sulfur cluster binding domain (PF00111) (3)
  - 3'-5' exonuclease family, domain 1 (PF01138) (3)
  - 3'-5' exonuclease (PF01612) (6)

41

You should obtain a results set something like that below. Does the set contain the orthologues for PfHT in *P. berghei* and *P. chabaudi*?



# Search results



Go To

Organisms

Go To

Shortcuts

[Help](#)

All of 11 results shown

				<a href="#">Report Download</a>
<i>P. falciparum</i>	CDS	<a href="#">PEI0785c</a>	transporter, putative	
<i>P. falciparum</i>	CDS	<a href="#">PEB0275w</a>	hypothetical protein, conserved	
<i>P. chabaudi</i>	CDS	<a href="#">PC000604.03.0</a>	sugar transporter, putative	
<i>P. chabaudi</i>	CDS	<a href="#">PC000855.00.0</a>	transporter, putative	
<i>P. chabaudi</i>	CDS	<a href="#">PC000736.00.0</a>	monosaccharide transporter, putative	
<i>P. berghei</i>	CDS	<a href="#">PB001031.03.0</a>	conserved hypothetical protein	
<i>P. falciparum</i>	CDS	<a href="#">PEB0210c</a>	monosaccharide transporter, putative	
<i>P. chabaudi</i>	CDS	<a href="#">PC000438.02.0</a>	conserved hypothetical protein	
<i>P. berghei</i>	CDS	<a href="#">PB000562.01.0</a>	monosaccharide transporter, putative	
<i>P. falciparum</i>	CDS	<a href="#">PEI0955w</a>	sugar transporter, putative	
<i>P. berghei</i>	CDS	<a href="#">PB000243.00.0</a>	transporter, putative	

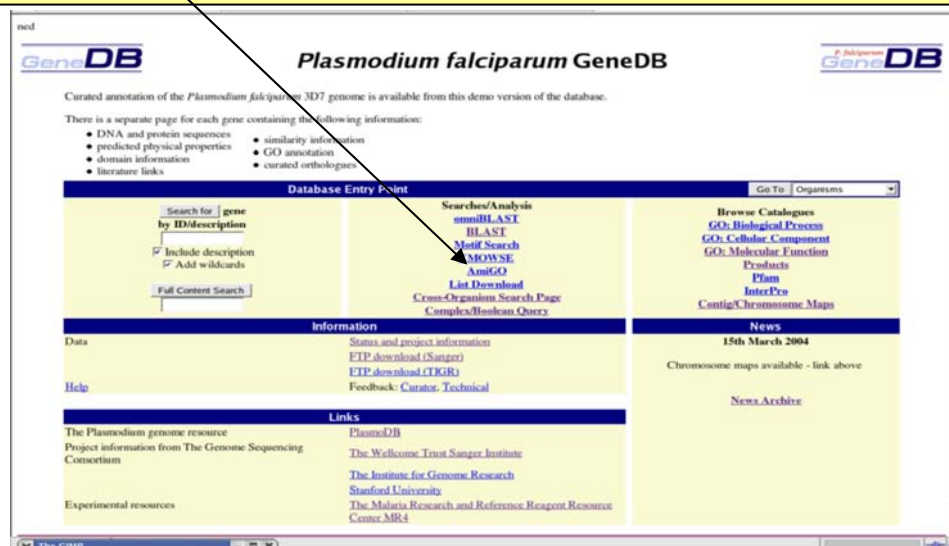
Hosted by the [Sanger Institute](#)

[Send comments, requests, corrections and updates](#)

## Exercise 4 Searching using Gene Ontology annotation

- 42 Another search strategy is to search on the basis of Gene Ontology terms. Gene Ontologies are structured vocabularies that are designed to describe biological processes in an accurate and consistent way (for more information see <http://www.geneontology.org>.) The ontology is composed of three terms: the molecular function, biological process and cellular component (location) of a protein. Where evidence exists from the literature, from sequence analysis or other sources Gene Ontology terms for function, process and component are attributed to that gene. AmiGO is a database of Gene Ontology associations that is designed and maintained by the Gene Ontology consortium. It allows searching and browsing of gene ontology annotation across many genomes (including those which are not annotated and curated for GeneDB) and is accessible via GeneDB. It can be a powerful way to search for genes with similar function across several organisms: in our case the search for transporters of glucose and hexose. The example below shows how to set up this query. Once you've tried it and have become familiar with it, try some of the other suggested searches or perhaps one that would be of interest to your own research.

- 43 Click on the AmiGO link to go to the AmiGO simple query search page.



- 44 Enter sugar transporter in the Search GO box (underlined red). In the datasource box (circled red) scroll down and click on GeneDB\_Pfalciparum, GeneDB\_Pbergei and GeneDB\_Pchabaudi (you'll need to press the shift button whilst clicking to select all three). Then click on the submit button (boxed red).





45

The results should be similar to those below – how do the results compare to your previous searches? Is this search successful in identifying the putative orthologues of PfHT in *P. chabaudi* and *P. berghei*? Clicking on the gene name (red arrow 1) will display the exact GO ontology annotation for that gene.

To go the GeneDB gene page click on the name of the database e.g. GeneDB\_Pfalciparum (red arrow 2).

The Term Lineage section shows that “carbohydrate transporter” term is a subset or “child” of transporter activity. Clicking on *GO:0005125: transporter activity* (arrow 3) will display all genes with transporter activity within *P. falciparum*, *P. berghei* and *P. chabaudi*.

The Associated Genes section can be used to apply the same search to other organism databases (arrow 4) and also to filter the results by evidence code (arrow 5). The evidence code provides information on the type of data that was used to apply a particular GO term to that gene. ISS is Inferred from Sequence or Structural similarity and is used when similarities such as BLAST hits, the presence of protein domains or other features based on sequence or structural similarity. IEA is inferred from Electronic Annotation and is used when similarities have been transferred from automated annotation and have not been reviewed by a curator. For a more detailed description of evidence codes click on the evidence link (arrow 6). If the evidence code has a link this will provide more information about the evidence for the GO term (arrow 7).

**carbohydrate transporter activity**  
 Accession: GO:0015144  
 Synonyms: sugar transporter  
 Definition: Enables the directed movement of carbohydrate into, out of, within or between cells.

**Term Lineage** [Graph view](#)  
 GO:0003673 : Gene Ontology (6113)  
 GO:0003674 : molecular function (5354)  
 GO:0005215 : transporter activity (237) ← 3  
 GO:0015144 : carbohydrate transporter activity (10)

**External References**  
 GO (1)

**Associated Genes** [Filter by database:](#) All, FB, SGD ← 4  
[Filter by Evidence for Association:](#) All, Curator Approved, Inferred from Mutant Phenotype ← 5  
[Filter Associated Genes](#)

Page 1

Gene Symbol:	Datasource:	Evidence:	Full name:
<a href="#">GO:0015145 : monosaccharide transporter activity</a>			
<a href="#">PF0210c</a> ← 1	<a href="#">GeneDB_Pfalciparum</a> ← 2	ISS	monosaccharide transporter, putative
<a href="#">PB000562.01.0</a>	<a href="#">GeneDB_Pberghei</a>	IEA - unpublished	Not Available
<a href="#">PC000736.00.0</a>	<a href="#">GeneDB_Pchabaudi</a>	IEA - unpublished	Not Available
<a href="#">GO:0005459 : UDP-galactose transporter activity</a>			
<a href="#">PF11_0141</a>	<a href="#">GeneDB_Pfalciparum</a>	ISS	UDP-galactose transporter, putative
<a href="#">PB001660.02.0</a>	<a href="#">GeneDB_Pberghei</a>	IEA - unpublished	Not Available
<a href="#">PC000722.01.0</a>	<a href="#">GeneDB_Pchabaudi</a>	IEA - unpublished	Not Available
<a href="#">GO:0005351 : sugar porter activity</a>			
<a href="#">PFE1455w</a>	<a href="#">GeneDB_Pfalciparum</a>	ISS	sugar transporter, putative
<a href="#">GO:0008524 : glucose 6-phosphate:phosphate antiporter activity</a>			
<a href="#">PFE0410w</a>	<a href="#">GeneDB_Pfalciparum</a>	ISS	triose or hexose phosphate / phosphate translocator, putative
<a href="#">PB000956.00.0</a>	<a href="#">GeneDB_Pberghei</a>	IEA - unpublished	Not Available
<a href="#">PC000805.01.0</a>	<a href="#">GeneDB_Pchabaudi</a>	IEA - unpublished	Not Available

Previous Page Next Page First Page [All Gene Products](#)  
[Check/Uncheck All](#) [Get Detailed View](#)

[Submit GO term or definition request.](#)  
[Submit AmiGO bug report.](#)

Copyright The Gene Ontology Consortium. All rights reserved.

46

Hopefully these exercises have familiarised you with several strategies for data mining in GeneDB, and given you ideas how GeneDB could be applied to your own research area. If you have any further questions please ask a demonstrator, or after the course please address your queries to the GeneDB team who will be happy to help you. See box 60 and the figure below for details of email links.

## Exercise 5 Use of the Artemis Applet

- 47** We are now going to look at the use of the artemis applet. It will only be considered briefly since you have already covered the use of Artemis. It can be launched from within GeneDB and is a useful way of viewing the gene in the context of the genome. It is especially useful for visualising intergenic regions, promoters, 5' and 3' untranslated regions, intron-exon boundaries, as well as many other features.
- 48** GO to the gene page for PFB0210c. You can do this in many ways, but one is to go to the *Plasmodium falciparum* page in GeneDB (see box 3). Then enter **PFB0210c** in the *Search for gene by ID/description* box (see box 4). Then from the gene page click on the **Graphical display in artemis** link

**GeneDB CDS: PFB0210c**

Search for: [ ] Go To: Organisms [ ] Go To: Shortcuts [ ] Help [ ] Contact curator [ ]

Sequence and annotation provided by TIGR (This gene at TIGR)

**General Information** [Add to Basket](#) [View Basket](#)

Name: PFB0210c  
 Systematic Name: PFB0210c  
 Status: Status unknown  
 Product: monosaccharide transporter, putative (0 Others)  
 Type: CDS  
 Sequence: [DNA and Protein](#)

**Location**

Chromosome: 2  
 Chromosome Location: complement(205889..207403) Length: 1515 bp  
 Exons: complement(205889..207403) (Spliced length: 1515 bp)

**Context Map:**

[PFB0195c](#) [PFB0200c](#) [PFB0205c](#) **PFB0210c** [PFB0215c](#) [PFB0220c](#) [PFB0225c](#) [PFB0230c](#)

**Predicted Peptide Properties**

Mass	56.4 kDa	Amino acids	504
Isoelectric point	pH 8.7	Charge	11.0
Signal Peptide	Not found		
Transmembrane	12 probable transmembrane helices predicted for PFB0210c by TMHMM2.0 at aa 20-42, 79-101, 108-125, 130-152, 164-186, 206-228, 293-315, 330-352, 359-378.		

**GeneDB Range Download Page**

**General Information**

Systematic id: PFB0210c  
 Location: complement(205889..207403)  
 Organism: *P. falciparum*  
 Type: CDS  
 Contig: MAL2 (Length: 947102 bp)

**Range Options**

☒ Up/Down Stream  
 3': [10000] 5': [10000]  
☒ From/To  
 From: [195889] To: [217403]

**Download Types**

☒ Download (features and sequence in EMBL format)  
☐ Download (sequence in Fasta format)  
☒ Artemis Applet ([Help on Artemis](#))  
 Please note the Java applet doesn't work on all browsers/OS combinations. If it doesn't run for you please [install Artemis locally](#).

[Submit Query](#) [Reset](#)

**49** This window allows you to specify the region that is opened by the viewer. The default is 10 Kb upstream and 10 Kb downstream of the gene selected. This can be modified by the user.

You can also selected the region that you want to view using coordinates.

**50** For the purposes of this example we will use the default settings. Click on the submit query button

52

The Artemis Applet retains nearly all of the functions that it has if it is run locally. Refer to your notes from module on Artemis if necessary. There are far too many functions to describe them all here so we are going to look at a few which are relevant to our investigation.

The hexose transporter that we are looking (PfHT, systematic identifier PFB0210c) has been characterised biochemically. It is able to transport glucose and fructose down a chemiosmotic gradient as a classic uniporter. Some residues that define substrate specificity have been identified by mutagenesis experiments (Woodrow et al., 2000). If the Glutamine residue at position 169 is changed to Asparagine, a mutation denoted by (Q169N), the ability to transport fructose is abolished, but the ability to transport glucose is retained. This residue exist within a the 5 th predicted transmembrane helix.

We'll use the artimis applet to:

- Look at the annotation for pHT-1 (systematic identifier PFB0210c)
- View hydrophobicity/hydrophilicity plots for the protein
- Examine the amino acid sequence around position 169



Use these scroll bars to adjust the DNA and protein views

53

Click here to select a gene and press E (or ctrl E) to view the annotation for the gene.

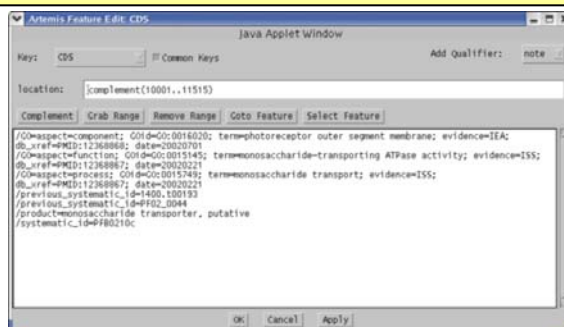
Find PFB0210C and view the annotation for this gene. It should be something like that shown below.

**Note the genes may be coloured due to updates in annotation**

54

Examine the annotation briefly before closing this window.

Select the gene of interest by **double** clicking on its entry in the gene list (green arrow) or on its box in the DNA or Protein view. A solid black line will appear around the gene in the DNA and protein view.

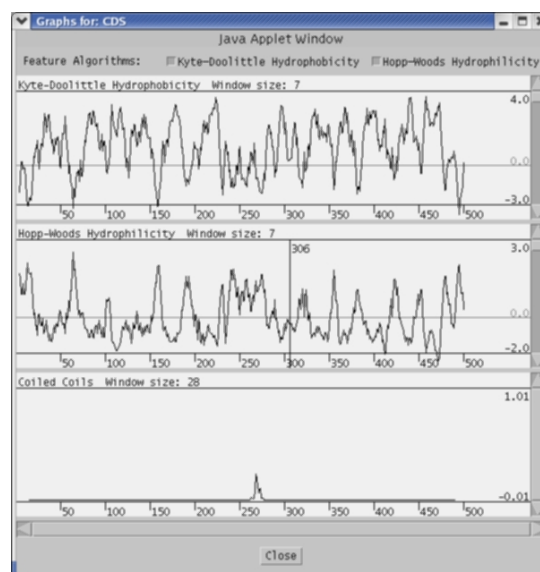


- 55 While the gene is select click on the *View* menu. Then from this drop down menu choose *show feature plots* (the bottom one in the list)



**Note: the genes in the display may appear coloured due to updates in annotation**

- 56 These plots show hydrophobicity (upper) and hydrophilicity (lower).  
Is residue 169 located within a hydrophilic or hydrophobic region of the protein? (note you can click within this diagram to get a line from the x-axis up to the curve)



- 57** Close this window. You should return to the Artemis window. Select the gene of interest again if it is not already selected.

58

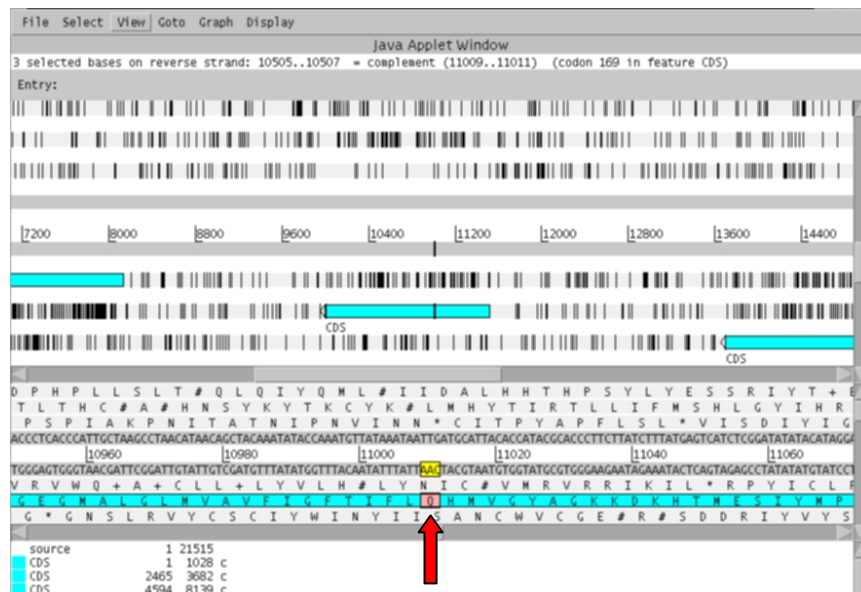
Now we are going to use the Artemis applet to find a specific amino acid, Glutamine Q at position 169, within the selected protein.

59

Click on *Goto* in the menu bar and select *Feature amino acid..*

A box will appear, enter 169 and click on OK.

The Glutamine residue will be highlighted in the protein view.



**60** If necessary use this scroll bar to adjust the protein view so that you can see individual amino acids

61

This part of the exercise has made you aware that you run use the Artemis applet from within GeneDB to view the detailed structure of the gene within its genomic content.

Annotation is at different stages for different genomes, and is actively improved where genes are manually curated. There will be occasions where the annotation may be misleading, incomplete or not as comprehensive as it could be. If you have any comments or about a particular gene's annotation, or can provide data that you think improves the knowledge base, then don't hesitate to contact the curator for that organism via email. Responses are normally provided within one or two working days. If your query or suggestion is of a technical nature, or is something that could apply to the whole of GeneDB, rather than the annotation of a particular gene or organism, then it may be better addressed to technical feedback. There are links on the bottom of each page in GeneDB (see image below).



# References

Rice,P. Longden,I. and Bleasby,A. **(2000)** *Trends in Genetics* **16** (6) 276-277  
EMBOSS: The European Molecular Biology Open Software Suite

Carver T.J., Mullan L.J., **(2002)** *Comparative and Functional Genomics* **3** (1) 75-78,  
A new graphical user interface to EMBOSS

Rutherford *et al.* **(2000)** *Bioinformatics* **16** (10) 944-945  
Artemis: sequence visualization and annotation

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.-A., Barrell, B.G. and Parkhill, J.  
**(2005)** *Bioinformatics* **21** (16) 3422-3423 ACT: the Artemis comparison tool.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape. (1997) Pathogenicity islands of virulent bacteria:structure, function and impact on microbial evolution. *Mol Microbiol* **23**;; 1089-97.

# Appendices

## **Appendix I: Artemis minimum hardware and software requirements.**

Artemis and ACT will, in general, work well on any standard modern machine and with most common operating systems. It is currently used on many different varieties of UNIX and Linux systems as well as Apple Macintosh and Microsoft Windows systems.

Note that the ability to run external programs (such as BLAST and FASTA) from within Artemis and ACT is available only on UNIX and Linux systems. Minimum memory requirements for people working on whole genomes are approximately 128 megabytes for Artemis and 128 megabytes per genome for ACT. Analysis of cosmid sized sequences can comfortably be achieved with less memory.

## **Appendix II: ACT comparison files**

ACT supports three different comparison file formats:

- 1) BLAST version 2.2.2 output: The blastall command must be run with the -m 8 flag which generates one line of information per HSP.
- 2) MEGABLAST output: ACT can also read the output of MEGABLAST, which is part of the NCBI blast distribution.
- 3) MSPcrunch output: MSPcrunch is program for UNIX and GNU/Linux systems which can post-process BLAST version 1 output into an easier to read format. ACT can only read MSPcrunch output with the -d flag.

Here is an example of an ACT readable comparison file generated by MSPcrunch -d.

```
1399 97.00 940 2539 sequence1.dna 1 1596 AF140550.seq
1033 93.00 9041 10501 sequence1.dna 9420 10880 AF140550.seq
828 95.00 6823 7890 sequence1.dna 7211 8276 AF140550.seq
773 94.00 2837 3841 sequence1.dna 2338 3342 AF140550.seq
```

The columns have the following meanings (in order): score, percent identity, match start in the query sequence, match end in the query sequence, query sequence name, subject sequence start, subject sequence end, subject sequence name.

The columns should be separated by single spaces.

### **Appendix III: Feature Keys and Qualifiers – a brief explanation of what they are and a sample of the one's we use.**

1 – Feature Keys: They describe features with DNA coordinates and once marked, they all appear in the Artemis main window. The ones we use are:

- ➔ CDS: Marks the extent of the coding sequence.
- ➔ RBS: Ribosomal binding site
- ➔ misc\_feature: Miscellaneous feature in the DNA
- ➔ rRNA: Ribosomal RNA
- ➔ repeat\_region
- ➔ repeat\_unit
- ➔ stem\_loop
- ➔ tRNA: Transfer RNA

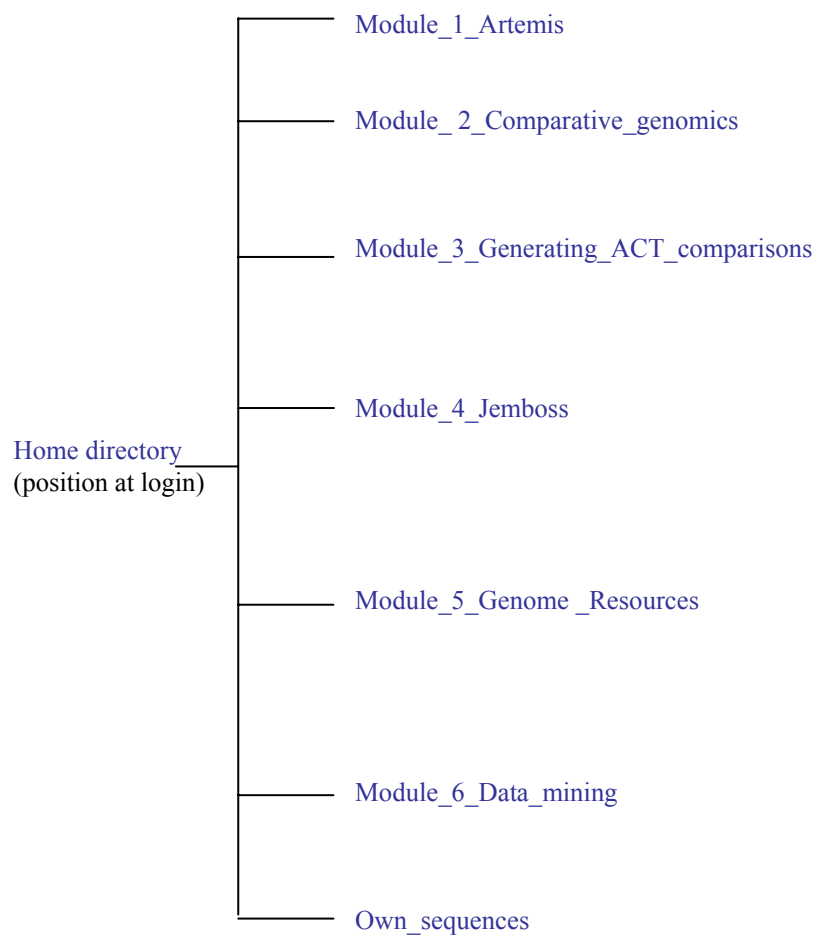
2 – Qualifiers: They describe features with protein coordinates. Once marked they appear in the lower part of the Artemis window. They describe the gene whose coordinates appear in the 'location' part of the editing window. The ones we commonly use for annotation at the Sanger Institute are:

- ➔ Class: Classification scheme we use “in-house” developed from Monica Riley's MultiFun assignments (see Appendix VI).
- ➔ Colour: Also used in-house in order to differentiate between different types of genes and other features.
- ➔ Gene: This qualifier either gives the gene a name or a systematic gene number.
- ➔ Label: Allows you to label a gene/feature in the main view panel.
- ➔ Note: This qualifier allows for the inclusion of free text. This could be a description of the evidence supporting the functional prediction or other notable features/information which cannot be described using other qualifiers.
- ➔ Partial: When a region in the DNA hits a protein in the database but lacks start and/or stop codons and the match does not include the whole length of the protein, it can be considered as a partial gene.
- ➔ Product: The assigned possible function for the protein goes here.
- ➔ Pseudo: Matches in different frames to consecutive segments of the same protein in the databases can be linked or joined as one and edited in one window. They are marked as pseudogenes. They are normally not functional and are considered to have been mutated.

The list of keys and qualifiers accepted by EMBL in sequence/annotation submission files are list at the following web page:

<http://www3.ebi.ac.uk/Services/WebFeat/>

## Appendix IV: Schematic of workshop files and directories

**Key:****Directories and subdirectories**



## Appendix V: Useful Web addresses

### Major Public Sequence Repositories

DNA Data Bank of Japan (DDBJ)	<a href="http://www.ddbj.nig.ac.jp">http://www.ddbj.nig.ac.jp</a>
EMBL Nucleotide Sequence Database	<a href="http://www.ebi.ac.uk/embl.html">http://www.ebi.ac.uk/embl.html</a>
Genomes at the EBI	<a href="http://www.ebi.ac.uk/genomes/">http://www.ebi.ac.uk/genomes/</a>
GenBank	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>

### Microbial Genome Databases Resources

Sanger Microbial Genomes	<a href="http://www.sanger.ac.uk/Projects/Microbes/">http://www.sanger.ac.uk/Projects/Microbes/</a>
TIGR Microbial Database	<a href="http://www.tigr.org/tdb/mdb/mdbcomplete.html">http://www.tigr.org/tdb/mdb/mdbcomplete.html</a>
Institute Pasteur GenoList databases <i>Including: SubtiList, Colbri, TubercuList, Leproma, PyloriGene, MypuList, ListiList, CandidaDB,</i>	<a href="http://genolist.pasteur.fr">http://genolist.pasteur.fr</a>
Pseudomonas Genome Database	<a href="http://www.pseudomonas.com/">http://www.pseudomonas.com/</a>
Clusters of Orthologous Groups of proteins (COGs)	<a href="http://www.ncbi.nlm.nih.gov/COG/">http://www.ncbi.nlm.nih.gov/COG/</a>
SCODBII ( <i>S. coelicolor</i> database)	<a href="http://www.jiio16.jic.bbsrc.ac.uk/S.coelicolor">http://www.jiio16.jic.bbsrc.ac.uk/S.coelicolor</a>

### Protein Motif Databases

Prosite	<a href="http://www.expasy.ch/prosite/">http://www.expasy.ch/prosite/</a>
Pfam	<a href="http://www.sanger.ac.uk/Software/Pfam/index.shtml">http://www.sanger.ac.uk/Software/Pfam/index.shtml</a>
BLOCKS	<a href="http://blocks.fhcrc.org">http://blocks.fhcrc.org</a>
InterPro	<a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a>
PRINTS	<a href="http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/">http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/</a>
SMART	<a href="http://smart.embl-heidelberg.de">http://smart.embl-heidelberg.de</a>
InterPro	<a href="http://www.ebi.ac.uk/interpro/index.html">http://www.ebi.ac.uk/interpro/index.html</a>

### Protein feature prediction tools

TMHMM Prediction of transmembrane helices in proteins	<a href="http://www.cbs.dtu.dk/services/TMHMM-2.0/">http://www.cbs.dtu.dk/services/TMHMM-2.0/</a>
SignalP Prediction Server	<a href="http://www.cbs.dtu.dk/services/SignalP/">http://www.cbs.dtu.dk/services/SignalP/</a>
PSORT protein prediction	<a href="http://psort.ims.u-tokyo.ac.jp/form.html">http://psort.ims.u-tokyo.ac.jp/form.html</a>

### Metabolic Pathways and Cellular Regulation

EcoCyc	<a href="http://ecocyc.org/">http://ecocyc.org/</a>
ENZYME	<a href="http://www.expasy.ch/enzyme/">http://www.expasy.ch/enzyme/</a>
Kyoto Encyclopedia of Genes and Genomes (KEGG)	<a href="http://www.genome.ad.jp/kegg">http://www.genome.ad.jp/kegg</a>
MetaCyc	<a href="http://ecocyc.org/">http://ecocyc.org/</a>

### Miscellaneous sites

NCBI BLAST website	<a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a>
The tmRNA website	<a href="http://www.indiana.edu/~tmrna/">http://www.indiana.edu/~tmrna/</a>
tRNAscan-SE Search Server	<a href="http://www.genetics.wustl.edu/eddy/tRNAscan-SE/">http://www.genetics.wustl.edu/eddy/tRNAscan-SE/</a>
Codon usage database	<a href="http://www.kazusa.or.jp/codon/">http://www.kazusa.or.jp/codon/</a>
RNAgenie RNA gene prediction	<a href="http://rnagene.lbl.gov/">http://rnagene.lbl.gov/</a>
GO Gene Ontology Consortium	<a href="http://www.geneontology.org/">http://www.geneontology.org/</a>
Artemis homepage	<a href="http://www.sanger.ac.uk/Software/Artemis/">http://www.sanger.ac.uk/Software/Artemis/</a>
ACT homepage	<a href="http://www.sanger.ac.uk/Software/ACT/">http://www.sanger.ac.uk/Software/ACT/</a>
Glimmer	<a href="http://www.tigr.org/software/glimmer/">http://www.tigr.org/software/glimmer/</a>
Orpheus	<a href="http://pedant.gsf.de/orpheus">http://pedant.gsf.de/orpheus</a>

## Appendix VI: Prokaryotic Protein Classification Scheme used within the PSU

This scheme was adapted for in-house use from the Monica Riley's protein classification

<<http://genprotec.mbl.edu/riley-lab.html>>).

More classes can be added depending on the microorganism that is being annotated (e.g secondary metabolites, sigma factors (ECF or non-ECF), etc).

- 0.0.0 Unknown function, no known homologs
- 0.0.1 Conserved in *Escherichia coli*
- 0.0.2 Conserved in organism other than *Escherichia coli*
- 1.0.0 Cell processes
  - 1.1.1 Chemotaxis and mobility
  - 1.2.1 Chromosome replication
  - 1.3.1 Chaperones
  - 1.4.0 Protection responses
    - 1.4.1 Cell killing
    - 1.4.2 Detoxification
    - 1.4.3 Drug/analog sensitivity
    - 1.4.4 Radiation sensitivity
  - 1.5.0 Transport/binding proteins
    - 1.5.1 Amino acids and amines
    - 1.5.2 Cations
    - 1.5.3 Carbohydrates, organic acids and alcohols
    - 1.5.4 Anions
    - 1.5.5 Other
  - 1.6.0 Adaptation
    - 1.6.1 Adaptations, atypical conditions
    - 1.6.2 Osmotic adaptation
    - 1.6.3 Fe storage
  - 1.7.1 Cell division
- 2.0.0 Macromolecule metabolism
- 2.1.0 Macromolecule degradation
  - 2.1.1 Degradation of DNA
  - 2.1.2 Degradation of RNA
  - 2.1.3 Degradation of polysaccharides
  - 2.1.4 Degradation of proteins, peptides, glycoproteins
- 2.2.0 Macromolecule synthesis, modification
  - 2.2.01 Amino acyl tRNA synthesis; tRNA modification
  - 2.2.02 Basic proteins - synthesis, modification
  - 2.2.03 DNA - replication, repair, restriction./modification
  - 2.2.04 Glycoprotein
  - 2.2.05 Lipopolysaccharide
  - 2.2.06 Lipoprotein
  - 2.2.07 Phospholipids
  - 2.2.08 Polysaccharides - (cytoplasmic)
  - 2.2.09 Protein modification
  - 2.2.10 Proteins - translation and modification
  - 2.2.11 RNA synthesis, modif., DNA transcrip.
  - 2.2.12 tRNA
- 3.0.0 Metabolism of small molecules
- 3.1.0 Amino acid biosynthesis
  - 3.1.01 Alanine
  - 3.1.02 Arginine
  - 3.1.03 Asparagine
  - 3.1.04 Aspartate
  - 3.1.05 Chorismate
  - 3.1.06 Cysteine
  - 3.1.07 Glutamate
  - 3.1.08 Glutamine
  - 3.1.09 Glycine
  - 3.1.10 Histidine
  - 3.1.11 Isoleucine
  - 3.1.12 Leucine
  - 3.1.13 Lysine
  - 3.1.14 Methionine
  - 3.1.15 Phenylalanine
  - 3.1.16 Proline
  - 3.1.17 Serine
  - 3.1.18 Threonine
  - 3.1.19 Tryptophan
  - 3.1.20 Tyrosine
  - 3.1.21 Valine
- 3.2.0 Biosynthesis of cofactors, carriers
  - 3.2.01 Acyl carrier protein (ACP)
  - 3.2.02 Biotin
  - 3.2.03 Cobalamin
  - 3.2.04 Enterochelin
  - 3.2.05 Folic acid
  - 3.2.06 Heme, porphyrin
  - 3.2.07 Lipoate
  - 3.2.08 Menaquinone, ubiquinone
  - 3.2.09 Molybdopterin
  - 3.2.10 Pantothenate
  - 3.2.11 Pyridine nucleotide
  - 3.2.12 Pyridoxine
  - 3.2.13 Riboflavin
  - 3.2.14 Thiamin
  - 3.2.15 Thioredoxin, glutaredoxin, glutathione
  - 3.2.16 biotin carboxyl carrier protein (BCCP)

**Appendix VI (cont):**

- 3.3.0 Central intermediary metabolism
  - 3.3.01 2'-Deoxyribonucleotide metabolism
  - 3.3.02 Amino sugars
  - 3.3.03 Entner-Doudoroff
  - 3.3.04 Gluconeogenesis
  - 3.3.05 Glyoxylate bypass
  - 3.3.06 Incorporation metal ions
  - 3.3.07 Misc. glucose metabolism
  - 3.3.08 Misc. glycerol metabolism
  - 3.3.09 Non-oxidative branch, pentose pathway
  - 3.3.10 Nucleotide hydrolysis
  - 3.3.00 other
  - 3.3.11 Nucleotide interconversions
  - 3.3.12 Oligosaccharides
  - 3.3.13 Phosphorus compounds
  - 3.3.14 Polyamine biosynthesis
  - 3.3.15 Pool, multipurpose conversions of intermed. metabol'm
  - 3.3.16 S-adenosyl methionine
  - 3.3.17 Salvage of nucleosides and nucleotides
  - 3.3.18 Sugar-nucleotide biosynthesis, conversions
  - 3.3.19 Sulfur metabolism
  - 3.3.20 Amino acids
- 3.4.0 Degradation of small molecules
  - 3.4.1 Amines
  - 3.4.2 Amino acids
  - 3.4.3 Carbon compounds
  - 3.4.4 Fatty acids
  - 3.4.5 Other
  - 3.4.0 ATP-proton motive force
- 3.5.0 Energy metabolism, carbon
  - 3.5.1 Aerobic respiration
  - 3.5.2 Anaerobic respiration
  - 3.5.3 Electron transport
  - 3.5.4 Fermentation
  - 3.5.5 Glycolysis
  - 3.5.6 Oxidative branch, pentose pathway
  - 3.5.7 Pyruvate dehydrogenase
  - 3.5.8 TCA cycle
- 3.6.0 Fatty acid biosynthesis
  - 3.6.1 Fatty acid and phosphatidic acid biosynthesis
- 3.7.0 Nucleotide biosynthesis
  - 3.7.1 Purine ribonucleotide biosynthesis
  - 3.7.2 Pyrimidine ribonucleotide biosynthesis
- 4.0.0 Cell envelop
  - 4.1.0 Periplasmic/exported/lipoproteins
  - 4.1.1 Inner membrane
  - 4.1.2 Murein sacculus, peptidoglycan
  - 4.1.3 Outer membrane constituents
  - 4.1.4 Surface polysaccharides & antigens
  - 4.1.5 Surface structures
- 4.2.0 Ribosome constituents
  - 4.2.1 Ribosomal and stable RNAs
  - 4.2.2 Ribosomal proteins - synthesis, modification
  - 4.2.3 Ribosomes - maturation and modification
- 5.0.0 Extrachromosomal
  - 5.1.0 Laterally acquired elements
    - 5.1.1 Colicin-related functions
    - 5.1.2 Phage-related functions and prophages
    - 5.1.3 Plasmid-related functions
    - 5.1.4 Transposon-related functions
- 6.0.0 Global functions
  - 6.1.1 Global regulatory functions
- 7.0.0 Not classified (included putative assignments)
  - 7.1.1 DNA sites, no gene product
  - 7.2.1 Cryptic genes

**Appendix VII: List of colour codes**

- 0** (white) - Pathogenicity/Adaptation/Chaperones
- 1** (dark grey) - energy metabolism (glycolysis, electron transport etc.)
- 2** (red) - Information transfer (transcription/translation + DNA/RNA modification)
- 3** (dark green) - Surface (IM, OM, secreted, surface structures)
- 4** (dark blue) - Stable RNA
- 5** (Sky blue) - Degradation of large molecules
- 6** (dark pink) - Degradation of small molecules
- 7** (yellow) - Central/intermediary/miscellaneous metabolism
- 8** (light green) - Unknown
- 9** (light blue) - Regulators
- 10** (orange) - Conserved hypo
- 11** (brown) - Pseudogenes and partial genes (remnants)
- 12** (light pink) - Phage/IS elements
- 13** (light grey) - Some misc. information e.g. Prosite, but no function

**Appendix VIII: List of degenerate nucleotide value/IUB Base Codes.**

**R = A or G**

**S = G or C**

**B = C, G or T**

**Y = C or T**

**W = A or T**

**D = A, G or T**

**K = G or T**

**N = A, C, G or T**

**H = A, C or T**

**M = A or C**

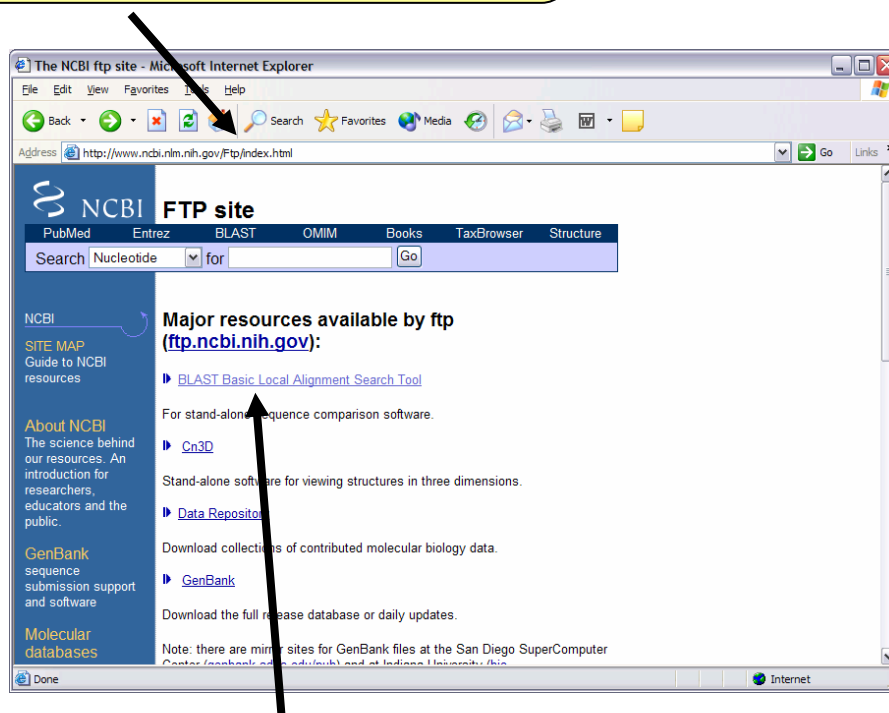
**V = A, C or G**

## Appendix IX: Downloading and installing BLAST on a Windows PC

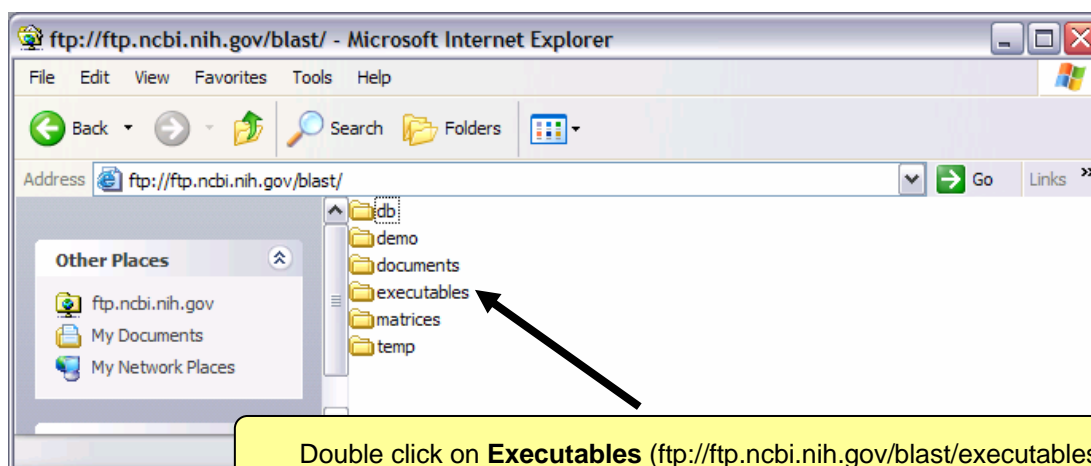
The following pages describe downloading BLAST onto a computer running Windows XP. Downloading onto computers with other versions of Windows should be essentially the same but the windows will look different to the screen shots used here.

Go to NCBI home page (<http://www.ncbi.nlm.nih.gov/>)

Scroll to bottom, Click on **FTP Site** (left hand side of the screen; <http://www.ncbi.nlm.nih.gov/Ftp/index.html>)

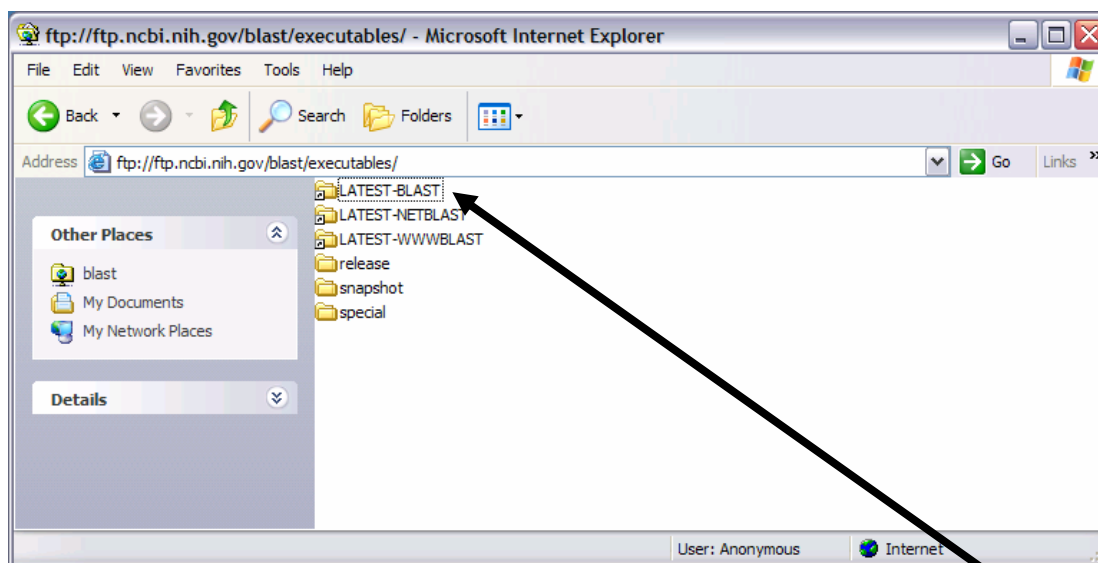


Click on **BLAST Basic Local Alignment Search Tool** (<ftp://ftp.ncbi.nih.gov/blast/>)

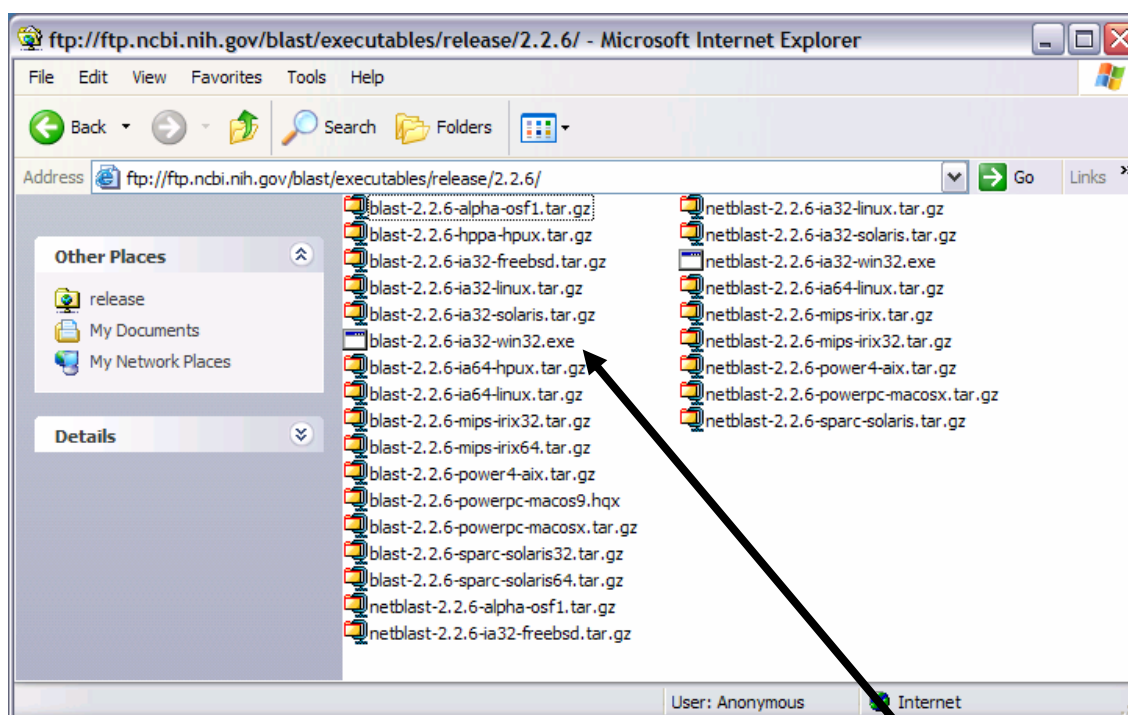


Double click on **Executables** (<ftp://ftp.ncbi.nih.gov/blast/executables/>)

This page may appear slightly different if you are using Netscape



Double click on the **LATEST-BLAST** shortcut

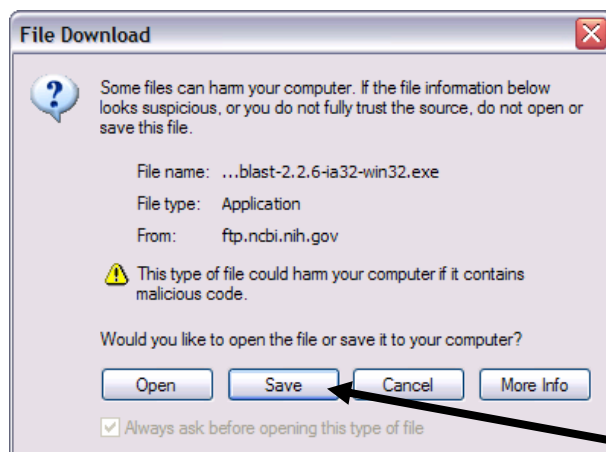


Double click on **blast-2.2.6-ia32-win32.exe**

Blast-2.2.6-ia32win32.exe is the blast exe file for windows



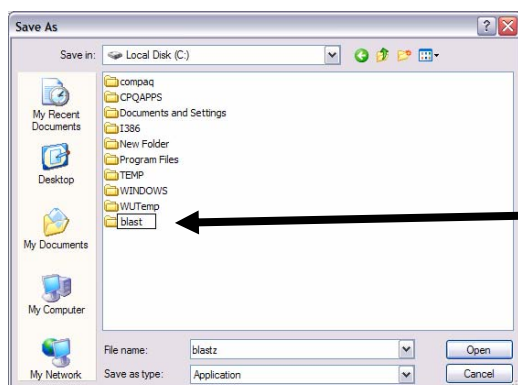
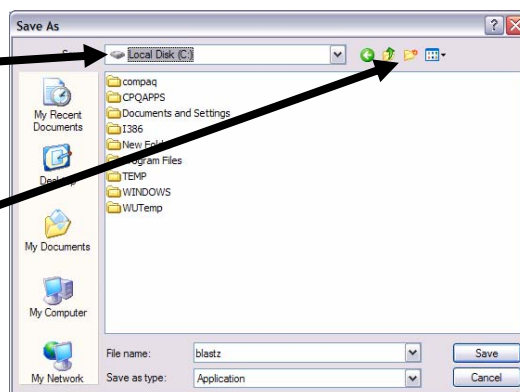
You now need to save the **blast-2.2.6-ia32-win32.exe** file in a new directory, blast, on the hard drive of your PC



Click on **Save**

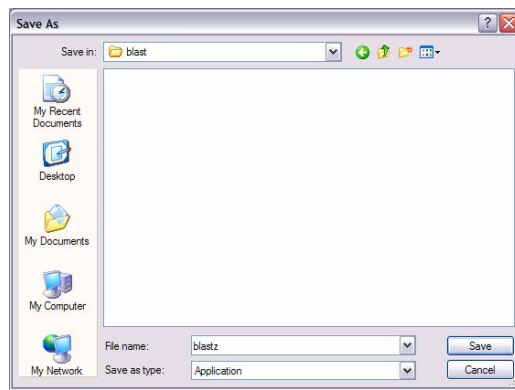
Click on **local disk C:**

Click on **new directory icon**  
(folder with a sun peeking through)



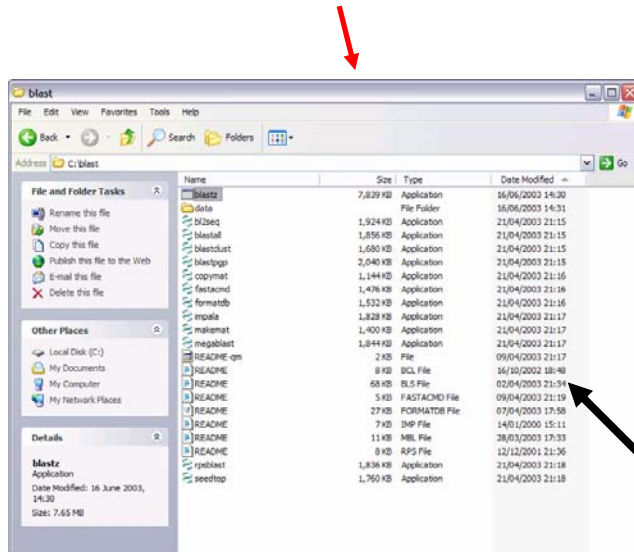
Type **blast** in the name box,  
press **Enter** key.

Double click on the new **blast**  
directory



Click on **Save**

Once downloaded view the contents of the blast directory by clicking on the open folder button



blast-2.2.6-ia32-win32.exe is a compressed file that contains a host of other files.

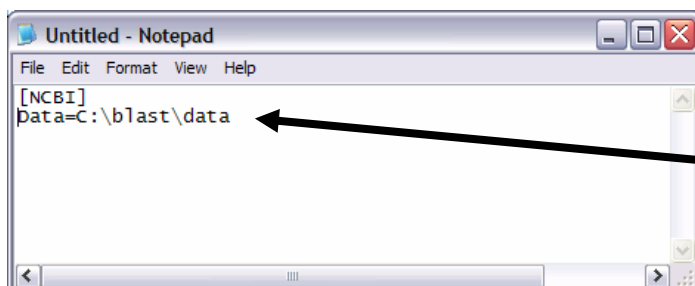
Now double click on the **blast-2.2.6-ia32-win32.exe** file to extract and unpack the rest of the BLAST download files

Included in the directory that has now been unpacked are several README files that describe the various programs in the BLAST software package. These files also provide descriptions of the command line options that you can set when you run the programs. To read these files double click on the icon or view them in notepad.

The **README.BLS** file contains details of the main BLAST program and how to format DNA sequences prior to running BLAST

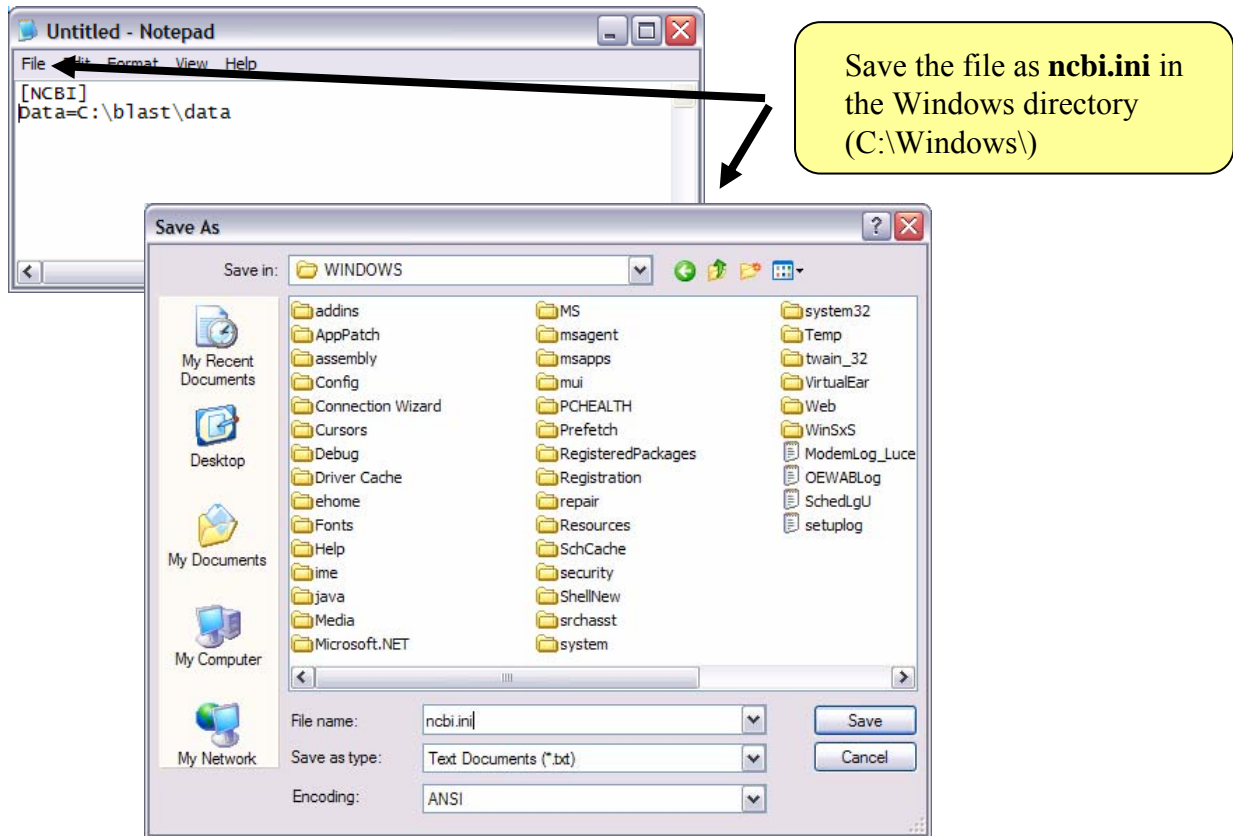
Before you can run BLAST you will need to create an **ncbi.ini** file containing the following lines:

```
[NCBI]
Data=C:\blast\data
```



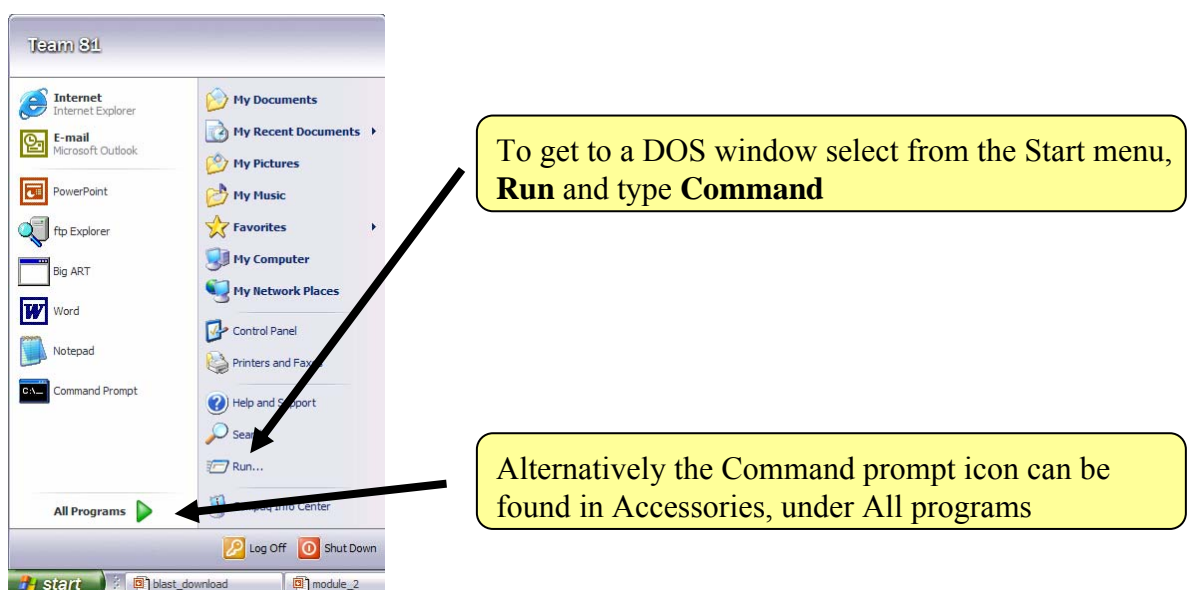
Open **Notepad** (All programs, Accessories menu). Type in the text:

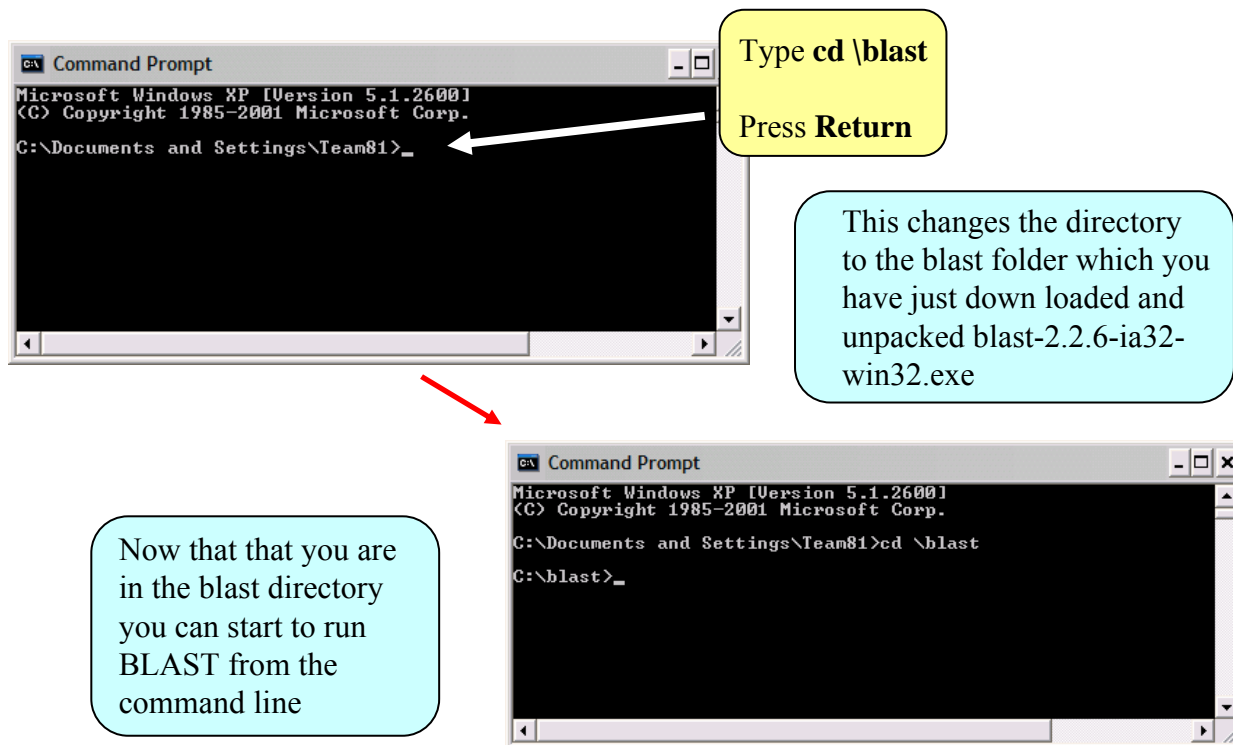
```
[NCBI]
Data=C:\blast\data
```



## Running BLAST

The BLAST software does not run in Windows, but DOS, an operating system that Windows runs in. When you want to run blast you will need a DOS window a.k.a. Command Prompt





There are several programs in the BLAST package that you have now downloaded that can be used for sequence comparison. For a detailed description of the uses and options see the appropriate README file.